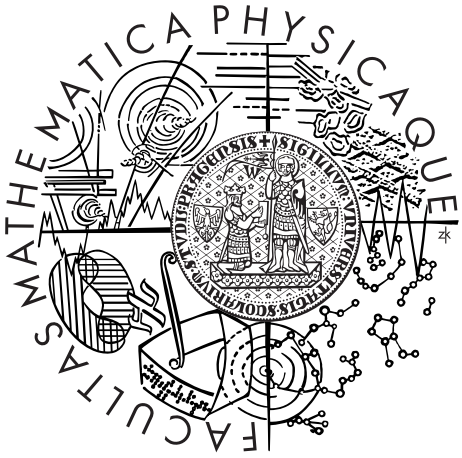
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Veronika Betíková

Role Business Intelligence a data-miningu v pojistném fraud managementu

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	Mgr. Pavel Pešout, Ph.D.
Studijní program:	Matematika
Studijní obor:	Finanční a pojistná matematika

Praha 2013

Na tomto mieste by som chcela poďakovať svojmu vedúcemu Mgr. Pavlovi Pešoutovi, Ph.D. za odborné vedenie tejto práce. Za ochotu a čas vždy pomôcť, jeho cenné rady, zaujímavé podnety, vecné pripomienky a za všetko, čo som sa pri písaní tejto práce naučila. Ďalej by som chcela poďakovať mojej rodine a priateľom za podporu počas môjho štúdia.

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne ..... Veronika Betíková

Název práce: Role Business Intelligence a data-miningu v pojistném fraud managementu

Autor: Veronika Betíková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Mgr. Pavel Pešout, Ph.D., Deloitte Advisory s.r.o.

Abstrakt: Tato práce se věnuje problematice pojistných podvodů a možnostem jejich řízení v rámci fraud managementu. V úvodu je pojistný podvod popsán obecně. Úloha odhalování pojistných podvodů pomocí prediktivních modelů je rozšířena o predikci profitability pojistného portfolia. Další části jsou zaměřeny na metody vícerozměrné statistiky, konkrétně na analýzu hlavních komponent, diskriminační analýzu a logistickou regresi. Tyto metody jsou popsány teoreticky a následně použity pro konstrukci prediktivních modelů. Uvedené metody jsou aplikovány na data z oblasti pojištění domácnosti. Závěrečná část práce pojednává o řízení fraud managementu pomocí Business Intelligence řešení. Navržen je systémový koncept fraud managementu, který má pomoci při odhalování pojistných podvodů.

Klíčová slova: pojistní podvod, Business Intelligence, analýza hlavních komponent, diskriminační analýza, logistická regrese

Title: The Role of Business Intelligence and Data Mining in the Insurance Fraud Management

Author: Veronia Betíková

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Pavel Pešout, Ph.D., Deloitte Advisory s.r.o.

Abstract: This thesis analyzes the problems of insurance fraud and the possibilities of their control in fraud management. Insurance fraud is described in general at the beginning. The role of insurance fraud detections using predictive models is extended to predict the profitability of insurance portfolio. Other parts are focused on methods of multivariate statistics, namely principal component analysis, discriminant analysis and logistic regression. These methods are described theoretically and then used in construction of predictive models. The methods are also applied on data from household insurance. The final part of the thesis discusses the control of the fraud management using Business Intelligence solution. There is designed concept of fraud management system, which should help in insurance fraud detection.

Keywords: insurance fraud, Business Intelligence, principal component analysis, dicriminant analysis, logistic regression

Názov práce: Rola Business Intelligence a data-miningu v poistnom fraud managemente

Autor: Veronika Betíková

Katedra: Katedra pravdepodobnosti a matematickej štatistiky

Vedúci diplomovej práce: Mgr. Pavel Pešout, Ph.D., Deloitte Advisory s.r.o.

Abstrakt: Táto práca sa venuje problematike poistných podvodov a možnostiam ich riadenia v rámci fraud managementu. V úvode je poistný podvod popísaný všeobecne. Úloha odhaľovania poistných podvodov pomocou prediktívnych modelov je rozšírená o predikciu profitability poistného portfólia. Ďalšie časti sú zamerané na metódy mnohorozmernej štatistiky, konkrétne na analýzu hlavných komponentov, diskriminačnú analýzu a logistickú regresiu. Tieto metódy sú popísané teoreticky a následne použité na konštrukciu prediktívnych modelov. Uvedené metódy sú aplikované na dáta z oblasti poistenia domácnosti. Záverečná časť práce pojednáva o riadení fraud managementu pomocou Business Intelligence riešenia. Navrhnutý je systémový koncept fraud managementu, ktorý má napomôcť pri odhaľovaní poistných podvodov.

Kľúčové slová: poistný podvod, Business Intelligence, analýza hlavných komponentov, diskriminačná analýza, logistická regresia

Obsah

Úvod

V posledných rokoch sme svedkami mnohých úsporných opatrení vo finančnom sektore, v oblasti poisťovníctva nevynímajúc. Nestále finančné výnosy, pokles prijatého poistného a neustávajúca snaha o udržanie konkurencieschopnosti na poistnom trhu núti poistiteľov šetriť a konsolidovať svoje náklady. Nie nadarmo sa hovorí, že najľahšie zarobené peniaze sú tie ušetrené. Jednou z oblastí, kde je v rámci poisťovne možné ušetriť nemalé finančné prostriedky, je zabrániť vyplácaniu poistných plnení, na ktoré nie je právny nárok. Poistné podvody sú významnou zložkou operačného rizika a ich efektívne riadenie môže ovplyvniť celkovú ziskovosť poisťovne.

Cieľom tejto práce je prispieť k zefektívneniu procesu odhaľovania poistných podvodov a navrhnúť riešenie riadenia fraud managementu s využitím Bussiness Intelligence a metód mnohorozmernej štatistky.

Úvodná kapitola je venovaná všeobecným poznatkom o poistnom podvode naprieč rôznymi odvetviami poistenia. Zhrnuté sú v nej posledné štatistky týkajúce sa poistných podvodov v Českej republike a taktiež postupy, ktorými sa poisťovne snažia chrániť pred neoprávnenými výplatami poistných plnení. My sa v našej práci zameriame na princíp odhaľovania poistných podvodov pomocou prediktívnych modelov.

Praktickú ilustráciu vybraných metód aplikujeme v práci na reálne dáta žiadostí o poistenie domácnosti. Našim cieľom pritom nebude predikovať len pravdepodobnosť výskytu poistného podvodu, ale úlohu rozšírime aj o odhad celkovej ziskovosti zmluvy. Našou predstavou je predložiť koncept, ako by poisťovne mohli filtrovať žiadosti o poistenie. A to takým spôsobom, aby uzatvárali zmluvy s vyššou pravdepodobnosťou pozitívneho škodného pomeru a tým zlepšovali profitabilitu svojho poistného kmeňa. V druhej kapitole je teda zhrnutá definícia tejto úlohy a popísané zdrojové dáta spolu so sledovanými znakmi.

V ďalších častiach práce sú potom postupne popísané vybrané metódy mnohorozmernej štatistiky, ktoré využijeme k tvorbe prediktívnych modelov. Výstavba jednotlivých modelov je v každej kapitole popísaná teoreticky a následne ilustrovaná na dátach z poistenia domácnosti. Použitý je pri tom softvér *NCSS 2007* a najmä *Wolfram Mathematica 8.0*, v ktorej sme úlohy programovali vytvorením vlastného kódu, ktorý taktiež uvádzame v prílohách tejto práce.

V tretej kapitole sa detailnejšie venujeme *analýze hlavných komponentov*. Túto metódu využijeme k prvej časti tvorby prediktívneho modelu, t. j. na identifikáciu premenných, ktoré sú podstatné pre našu úlohu z hľadiska zachovania čo najväčšej časti celkovej variability dátovej štruktúry. V praktickej časti úlohy je naším cieľom identifikovať tie merané znaky jednotlivých žiadostí, ktoré nemusia byť

pre účely predikcie profitability spracovávané v ďalších častiach práce.

Konštrukciou samotných prediktívnych modelov sa následne zaoberáme v štvrtej kapitole. Vychádzame pritom z poznatkov *diskriminačnej analýzy*. Pomocou kanonickej diskriminačnej analýzy je možné určiť významné premenné na odlíšenie zadefinovaných skupín pozorovaní, ktoré v úlohe poistenia domácnosti tvoria profitabilné a neprofitabilné zmluvy roztriedené podľa svojho dlhodobého škodného pomeru. Kanonickú diskriminačnú analýzu a rovnako tak metódu kvadratických diskriminačných funkcií ďalej využijeme na odhad profitability zohľadňujúcej poistný podvod pre sadu nových žiadostí.

Piata kapitola je venovaná tvorbe prediktívneho modelu pomocou *logistickej regresie*, ktorá sa v poisťovníctve využíva najmä na modelovanie odhadu pravdepodobnosti poistnej udalosti, či stornovanie poistnej zmluvy. Logistická regresia je taktiež využitá na klasifikáciu nových žiadostí a dosiahnuté výsledky sú porovnané s výsledkami získanými na základe diskriminačnej analýzy.

Záverečná kapitola je zameraná na prepojenie tvorby prediktívnych modelov s aplikačnými a systémovými riešeniami poisťovní. Navrhujeme v nej inovačný koncept fraud managementu systému ako integráciu jeho prvkov do dátovej štruktúry spoločnosti. Využité sú pritom zložky *Business Intelligence*. Predstavené sú tiež možnosti ako pomocou systémových riešení dosiahnuť čo najefektívnejšie spracovanie preditkívnych modelov. A taktiež ako zabezpečiť vhodné využitie výstupov z jednotlivých modelov v rámci oddelenia likvidácie poistných udalostí alebo pri prešetrovaní podozrivých jednaní.

Kapitola 1

Poistný podvod a jeho charakteristiky

1.1 Poistný podvod

V súčasnej dobe celosvetovej recesie sa téma poistných podvodov stáva z roka na rok aktuálnejšou. Zhoršujúca sa ekonomická situácia v domácnostiach nabáda ľudí čoraz častejšie k páchaniu poistných podvodov.

V posledných rokoch pozorujeme naprieč celým poistným trhom nárast poistných podvodov, väčšiu vynaliezavosť a zapájanie sa väčších organizovaných skupín do ich páchania. Fingované dopravné nehody, krádeže vozidiel, nadhodnocované výšky škôd pri prírodných katastrofách, úmyselne zakladané požiare nehnuteľností, ale i falošné lekárske správy, zatajovanie skutočného zdravotného stavu, či vymyslené úrazy patria k častým praktikám, pomocou ktorých sa poistení snažia obohatiť na úkor poistiteľov.

Hlavným motívom poistiteľov na odhaľovanie týchto podvodov je zabránenie finančným stratám a nevyplácanie plnení, na ktoré nie je právny nárok. Avšak snaha o riešenie tohto problému by nemala byť len na strane poistiteľov. Zapojiť do boja proti poistným podvodom by sa mali všetci, ktorí pravidelne platia poistné a nemajú žiadne postranné úmysly. Mnohí ľudia ani nevedia, že dôsledkom poistných podvodov je mnohokrát zvýšenie sadzieb poistného. Neodhalené poistné podvody totiž zhoršujú celkový škodný priebeh poistenia a tým pádom ovplyvňujú aj výšku poistného všetkých klientov. Teda náklady spôsobené poistným podvodom v konečnom dôsledku nesú všetci poistníci, i tí poctiví.

Poistný podvod možno vo všeobecnosti charakterizovať ako úmyselné klamanie jednej strany druhou v poistnom vzťahu za účelom obohatenia sa alebo získania výhody, ktorá by nenastala, ak by bol pravdivo vysvetlený skutkový stav.

Poistný podvod má i svoju právnu úpravu[1] v legislatíve Českej republiky. Jedná sa o trestný čin, ktorý sa zaraďuje do kategórie majetkovej kriminality.

[1]Právne predpisy súvisiace s poistným podvodom v ČR sú: *Trestní zákon, Zákon o pojišťovnictví, Zákon o pojistné smlouvě, Zákon o ochraně osobních údajů, Zákon o pojištení odpovědnosti z provozu vozidla, Zákon o pojišťovacích zprostředovatelích a likvidátorech pojistních událostí.*

Trestný zákonník[2] definuje *poistný podvod* ako dve samostatné skutkové podstaty:

1. Kto uvedie nepravdivé alebo hrubo skreslené údaje alebo podstatné údaje zamlčí:
   (a) v súvislosti s uzatváraním alebo zmenou poistnej zmluvy
   (b) v súvislosti s likvidáciou poistnej udalosti
   (c) pri uplatnení práva na plnenie z poistenia alebo iné obdobné plnenie bude potrestaný odňatím slobody až na 2 roky, zákazom činnosti alebo prepadnutím veci alebo inej majetkovej hodnoty.
2. Kto uvedie v úmysle opatriť sebe alebo inému prospech, vyvolá alebo predstiera udalosť, s ktorou je spojené právo na plnenie z poistenia alebo iné obdobné plnenie, alebo stav vyvolaný poistnou udalosťou udržuje a spôsobí tak na cudzom majetku nezanedbateľnú škodu, bude rovnako potrestaný.

Ďalej Trestný zákonník uvádza výšku jednotlivých trestov podľa rozsahu škody. Tresty sa pohybujú v rozmedzí 6 mesiacov až 10 rokov odňatia slobody. Presné znenie možno nájsť v [12].

1.2 Typy poistných podvodov

K poistnému podvodu môže dôjsť v ktorejkoľvek fáze poistnej zmluvy:

- pri zjednávaní a podpise poistnej zmluvy
- pri upisovaní a spravovaní poistnej zmluvy
- pri likvidácií poistných udalostí.

Pri zjednávaní poistnej zmluvy sa najčastejšie jedná o zatajenie dôležitých informácií pre ocenenie rizika. Z obavy, že poistiteľ odmietne poskytnúť poistnú ochranu alebo z dôvodu získania lepších podmienok. Patrí sem i uzatváranie zmlúv na neexistujúci majetok, či stanovovanie vyššej hodnoty majetku. V priebehu poistnej zmluvy sa poistný podvod môže objaviť pri dodatočnej zmene údajov, napríklad prehodnotenie zdravotného stavu na základe skreslených lekárskych správ. Zámerné spôsobenie škody, ktorú klient nahlási ako poistnú udalosť, prípadne nahlásenie vyššej škody k akej skutočne došlo, sú príkladmi podvodných jednaní pri likvidácií poistných udalostí.

Existuje mnoho kritérií podľa ktorých možno poistné podvody deliť. Jedným z nich je druh poistenia. Poistné podvody sa vyskytujú v rámci životného i neživotného poistenie a teda v:

- poistení majetku, vozidiel, zodpovednosti

[2] Vyhláška č. 40/2009 Sb. §210

- poistení osôb
- cestovnom poistení.

Páchateľom podvodu nemusí byť vždy len poistený. Podľa osoby páchateľa môžeme poistné podvody rozdeliť na **interné**, kde je páchateľom zamestnanec poisťovne, a **externé** kedy je páchateľom poistník, poškodený alebo poistený. K omnoho závažnejším podvodom, ktoré spôsobujú i vyššie škody, patria interné podvody. Pri týchto typoch podvodov je páchateľom zamestnanec poisťovne, ktorý využíva znalosti vnútornej štruktúry, systémov a procesov vo svoj prospech, resp. k osobnému obohateniu sa. V dnešnej dobe nie je výnimkou, že tieto typy podvodov nie sú páchané len z vlastnej iniciatívy, ale v spolupráci s páchateľmi mimo poisťovne.

1.3 Poistné podvody v číslach

CEA (Comité Européen des Assurances) odhaduje, že plnenie poistných podvodov predstavuje asi 2 % ročného predpísaného poistného, viď [8]. Na českom trhu je však situácia ešte vážnejšia. Odhady latencie poistných podvodov (celkový počet vrátane neodhalených) sa podľa [3] pohybujú medzi 5 - 15 % predpísaného poistného s ohľadom na druh poistenia. Znamená to teda, že české poisťovne tratia ročne okolo 8,5 až 12,5 miliárd Kč na výplatách za poistné podvody. Podľa svetových štatistík zhruba každý siedmy prípad, ktorý klient poisťovniam ohlási, patrí medzi poistné podvody. I to je jeden z dôvodov prečo ich nesmieme brať na ľahkú váhu.

Česká asociácia poisťovní (ČAP) každoročne spracováva štatistiku poistných podvodov, ktoré boli odhalené jej členskými poisťovňami. V tabuľke 1.1 je uvedený prehľad počtu prípadov vyšetrovania poistných podvodov v jednotlivých oboroch poistenia.

Obor poistenia	**Počet šetrených prípadov**					**Výška preukázaných hodnôt (v tis. Kč)**				
Roky	**2007**	**2008**	**2009**	**2010**	**2011**	**2007**	**2008**	**2009**	**2010**	**2011**
Poistenie vozidiel	3359	3510	3110	3211	4728	269593	347484	360072	302904	329730
Poistenie prepravy	15	11	28	21	17	4955	1875	3803	6353	1317
Poistenie majetku a zodpovednosti	654	595	817	967	891	216617	168375	237868	268517	425569
Poistenie osôb	520	690	523	943	1046	32803	32919	30672	47213	82461
Spolu	4584	4806	4478	5142	6682	523968	550653	632415	624987	839077

Tabuľka 1.1: Štatistiky poistných podvodov v ČR, roky 2007 - 2011

Experti z komerčných poisťovní v ČR sa zhodujú, že najviac pokusov o podvodné jednanie je i naďalej v poistení motorových vozidiel. Najčastejšie sú krádeže vozidiel, súčiastok, predmetov ponechaných vo vozidle a samozrejme fingované dopravné nehody. Výrazne však rastie i počet podvodov v poistení zodpovednosti.

Z tabuľky 1.2 je zjavné, že i v roku 2012 bolo najviac šetrených prípadov v poistení vozidiel. Za povšimnutie stojí najvyššia preukázaná hodnota v poistení majetku a zodpovednosti, ktorej podiel na všetkých podvodoch je približne 50%. Nárast pozorujeme i v poistení osôb. Zaujímavým je aj celkový pokles šetrených prípadov oproti roku 2011 a naopak nárast preukázanej hodnoty v poslednom sledovanom roku. Tento fakt naznačuje zvýšenú efektivitu poisťovní v odhaľovaní poistných podvodov.

Obor poistenia	**Počet šetrených prípadov**	**Výška nárokovaných plnení (v tis. Kč)**	**Výška preukázanej hodnoty (v tis. Kč)**
Poistenie vozidiel	3855	618120	371495
Poistenie prepravy	38	21748	19079
Poistenie majetku a zodpovednosti	954	805430	562831
Poistenie osôb	1296	122814	111171
Spolu	6143	1568112	1064576

Tabuľka 1.2: Štatistiky poistných podvodov v ČR, rok 2012

1.4 Boj s poistným podvodom

Dôvody narastajúcej aktivity boja proti poistným podvodom, ktoré potvrdzuje aj ČAP, sú celkom zrejmé. Súčasný trend stagnácie prijatého poistného a nestabilných, či nejasných výnosov z finančných operácií znamená evidentnú potrebu poisťovní zvyšovať svoj zisk, respektíve znižovať náklady, ku ktorým patria aj čiastky vyplácané za podvody. Znížením počtu úspešných podvodov, t. j. i škodového pomeru, si teda poisťovne môžu zaistiť udržanie svojej konkurencieschopnosti.

Nezanedbateľnou súčasťou boja proti poistným podvodom je efektívny všeobecný systém riadenia rizika. Tento systém zahŕňa aj **fraud management**, respektíve riadenie rizík podvodov, ktorý spadá do riadenia operačných rizík. Komplexný cyklus fraud managementu, pozri [13], je tvorený nasledujúcimi zložkami:

- Odstrašovanie (zabránenie podvodu predtým než sa oň človek vôbec pokúsi).

- Prevencia (predchádzanie výskytu podvodov). Táto fáza nastáva, ak zlyhajú všetky činnosti vo fáze odstrašovania.
- Detekcia (identifikácia a lokalizácia podvodov pred, v priebehu a po dokončení podvodnej činnosti; odhalenie prítomnosti podvodu alebo pokusu o podvod).
- Zmierňovanie (zastavovanie podvodu, keď nastane alebo zmiernenie straty z prebiehajúcich podvodov, napríklad zablokovaním účtu).
- Analýza (identifikácia podvodu, ktorý nastal aj napriek predchádzajúcim fázam, vyhodnotenie objemu a príčin strát z podvodu, dopadu na fraud management).
- Stratégia (vytvorenie zásad a stratégie, ktorá bude viesť k zníženiu výskytu podvodov).
- Vyšetrovanie (získanie informácií na zastavenie podvodnej činnosti, dokladanie dôkazov pre trestné stíhanie a odsúdenie podvodníkov).
- Trestné stíhanie, obžaloba (presadzovanie práva, obvinenie a stanovenie trestu pre páchateľa).

Jednotlivé fázy tohto cyklu sa mnohokrát prelínajú a bývajú zredukované na tri hlavné zložky: prevencia, detekcia a vyšetrovanie.

Zlepšenie fraud managementu si (i keď nepriamo) ako jeden zo svojich cieľov kladie aj pripravovaný regulatórny koncept Solventnosť II, ktorý by mal vstúpiť do platnosti 1. januára 2015. Jeho cieľom je poskytnúť vyššiu ochranu poisteným za súčasnej stability trhu. V koncepte sú okrem iného kladené nároky na efektívny systém riadenia operačného rizika, ktoré zahŕňa aj poistné podvody. Nároky sú kladené teda i na implementáciu procesu identifikácie a hodnotenie podvodu. Požiadavky zahŕňajú existenciu zdokumentovanej stratégie, rozsahu riadenia a taktiež procesy posudzovania a reportovania.

Keďže operačné riziko je súčasťou formule pre výpočet solventnostného kapitálového požiadavku i efektívnym riadením fraud managementu bude možné znižovať jeho hodnotu. A to môže byť v dobe, kedy by podľa dopadovej štúdie Solventnosti II (QIS 5 - The fifth Quantitative Impact Study, viď [11]) minimálne 15 % poisťovní v Európskej Únií nesplnilo požiadavky na solventnostný kapitálový požiadavok, významnou motiváciou.

Napriek tomu, že poisťovne v poslednej dobe vykazujú vyššiu aktivitu v riadení procesov svojho fraud managementu, nie je situácia zďaleka optimálna. V Českej republike sa poisťovniam podarilo v roku 2012 odhaliť poistné podvody za približne 1 miliardu Kč. Táto suma však tvorí len 6 -9 % predpokladaných výplat za podvody. Prečo je toto číslo také malé?

Súčasná situácia v riadení poistných podvodov je zložitá. Boj proti nim v poslednej dobe komplikujú aj meniace sa podmienky na trhu, vrátane niektorých pre klientských prístupov poisťovní, ako sú moderné formy zjednávania poistení, či hlásenia škôd cez internetové rozhrania. Aj napriek tomu možno základné dôvody

neuspokojivého stavu fraud managementu zvyčajne hľadať niekde inde. Základnými dôvodmi sú najmä:

- chýbajúca stratégia a nejednotnosť v prístupe v boji proti podvodom
- nefunkčné systémy na odhaľovanie podvodov a zlá kontinuita jednotlivých procesov
- nedostatočná motivácia zodpovedných zamestnancov
- protichodné ciele obchodu a riadenia rizík, či zlá spolupráca medzi jednotlivými oddeleniami spoločnosti
- nedostatok vhodných dát
- chýbajúci software zameraný na odhaľovanie podvodov
- nesystémové definície rolí a zodpovedností.

Vzhľadom k týmto komplexným problémom si vyžaduje i smerovanie aktivít v boji proti podvodom komplexný prístup. Dôraz musí byť kladený na centralizáciu procesov a jednotný prístup riadenia rizík v spoločnosti, fraud skóring a implementáciu systémov na odhaľovanie podvodov. Nemenej dôležité je aj budovanie povedomia o poistných podvodoch a ich prevencia. Vedomie, že poisťovňa sa podvodmi zaoberá odradí mnoho potencionálnych podvodníkov z obavy možného odhalenia.

Medzi pozitívne trendy, ktoré sa v poslednom období objavujú, patrí efektívne rozdelenie činností a reorganizácia v rámci poisťovní. Ako príklad sa často uvádza presúvanie vyšetrovania menej závažných podvodov na oddelenie likvidácií škôd. Tým pádom majú špecializované oddelenia viac priestoru na odhaľovanie závažnejších podvodov, ktorých preukazovanie je odborne i časovo náročnejšie. Ďalším z trendov je investícia do špeciálnych nástrojov na odhalenie poistných podvodov.

Dôležitou súčasťou boja proti poistným podvodom je vzájomná spolupráca poistiteľov. V rámci Českej republiky prebieha prostredníctvom *Českej kancelárie poistiteľov* (ČKP), či ČAP. Ide najmä o výmenu informácií o preverovaných poistných udalostiach, jednotlivých šetreniach a prípravu preventívnych opatrení. Na jar roku 2012 ČAP spustila nový systém pre odhaľovanie poistných podvodov SVIPO t. j. *Systém výmeny informácií o podozrivých okolnostiach.* Jeho cieľom je systematicky odhaľovať podvody, ktoré sú páchané na trhu a doplniť už zaužívané metódy. Tento systém porovnáva poistné udalosti vyšetrované jednotlivými poisťovňami a upozorňuje na tie, ktoré sú s veľkou pravdepodobnosťou poistnými podvodmi. Pripojiť sa k tomuto systému môžu všetky komerčné poisťovne pôsobiace na českom trhu. V súčasnosti systém pracuje s dátami z oblasti motorových vozidiel, avšak je plánované jeho rozšírenie do ďalších odvetví poistenia. Tento systém je určite vhodným doplňujúcim nástrojom v boji s poistnými podvodmi. Spracúva však už vyšetrované zmluvy a tým pádom v úlohách poisťovní i naďalej zostáva selektovanie podozrivých zmlúv, respektíve poistných udalostí.

Uvedomme si však, že ani tie najlepšie procesy, metodika, či štandardy riadenia samy o sebe nezaistia efektívne odhaľovanie podvodov. Hlavným aspektom je totiž vždy uplatnenie metód detekcie a vyšetrovania v kombinácii s ich integráciou v informačných systémoch danej spoločnosti. Tie sú často kľúčom k úspechu alebo zlyhaniu činností v jednotlivých fázach spomínaného cyklu a vedú k úspechu, či naopak zlyhaniu celého oddelenia fraud managementu.

V dnešnej dobe už existujú systémy, ktoré sú pomocou dátovej analýzy schopné poistné podvody odhaliť. Rozlíšiť možno dva základné princípy, na ktorých sú založené:

- princíp **expertných pravidiel**
- princíp **prediktívnych modelov**.

Cieľom expertných pravidiel je charakterizovať podozrivé udalosti. Tieto pravidlá sú vytvárané na základe expertných znalostí a skúseností. Aplikujú sa pri tom **indikátory potencionálnych podvodov**. Jedná sa o súhrn odpozorovaných a overených okolností priebehu škodnej udalosti v praxi, ktoré signalizujú, že ide o potencionálny poistný podvod. Pracovníci poisťovní využívajú tieto indikátory pri likvidácií, či revízii poistných udalostí na odhaľovanie poistných podvodov. K indikátorom poistných podvodov, ak hovoríme vo všeobecnej rovine, patria napríklad:

- viac škôd v prvých 90 dňoch od počiatku poistenia
- atypická škodná frekvencia a ich ročný nárast
- častá zmena poistiteľov
- absencia poistnej histórie
- veľké množstvo malých poistných udalostí
- chýbajúce lekárske potvrdenia, zmeny lekárov
- dokumentačné indikátory účty za ošetrenia nemajú obvyklé záležitosti, pozmenené dokumenty, sporná autentickosť dokladov.

Vedľa všeobecných pravidiel zameraných na atribúty jednotlivých zmlúv a ich životných cyklov sú expertné pravidlá používané poisťovňami aj pre detekciu fraudov z interného pohľadu. Tu sa prvotná kontrola neprevádza na zmluvách a klientoch, ale vychádza sa z parametrov sledovaných pre jednotlivé predajné kanály resp. sprostredkovateľov poistenia. V nasledujúcej časti sú zhrnuté niektoré sledované ukazovatele pre poisťovacích agentov a maklérov[3]. Poisťovňa na základe stanovených kritérií posudzuje, kedy sledované ukazovatele naznačujú možný po-

[3]Agent predáva produkty poisťovne, s ktorou má uzavretú zmluvu, na základe ktorej mu náleží provízia. Maklér naopak nepredáva produkty konkrétnej poisťovne, ale vyhľadáva pre klientov najvýhodnejšie produkty na trhu

istný podvod. Po prekročení interne stanovených úrovní nastáva hlbšie šetrenie vybraných sprostredkovateľov a ich klientov.

- **Ukazovatele pre agentov:**
  - Nový obchod jednotlivých agentov za daný mesiac. Pre existujúcich klientov sa zároveň sleduje stav ich portfólia.
  - Sledujú sa aktívni agenti s províznym dlhom.
  - Zmluvy za posledný rok uzatvorené na seba samého, či iného agenta.
  - Ukončenie zmluvy do 80 dní od uzatvorenia zmluvy.
  - Zmluvy, ktoré zanikli za posledný rok po maximálne 2 rokoch platenia poistného.
- **Ukazovatele pre maklérov:**
  - Skokový nárast produkcie nového obchodu, prehľad produkcie maklérov.
  - Podiel maklérov na dlhu maklérskej spoločnosti.
  - Kontrola maklérov s klientmi, ktorí majú anualizované poistné vyššie než určené čiastka.
  - Kontrola jednotlivých zmlúv s anualizovaným poistným vyšším než určená čiastka.
  - Sledovanie dlhu na firemných zmluvách, aby sa predišlo prípadnému zániku všetkých zmlúv vo firme.
- **Spoločné kritériá pre agentov a maklérov:**
  - Významný podiel zmlúv s vysokým poistným.
  - Poistený vystupuje na viacerých zmluvách.
  - Prehľad zmlúv, kde je uhradené maximálne tretie poistné, porovnanie s počtom nových zmlúv za dané obdobie.
  - Platenie poistného zloženkou.
  - Z jedného účtu hradené poistné pre viac zmlúv.
  - Aktuálny prehľad dlhov po agentoch/ maklérоch.
  - Percento zrušených zmlúv.
  - Prehľad provízneho dlhu agentského/maklérskeho kanálu po jednotlivých poradcoch.
  - Kontrola, či klientské platby neprichádzajú z provízneho účtu poradcu.

Treba podotknúť, že hoci nám tieto indikátory naznačia možný výskyt podvodu, je nutné aby si likvidátori zachovali objektivitu, keďže nie každý má v úmysle poistný podvod spáchať.

Ďalšími z techník detekcie podvodov sú spomínané prediktívne modely, ktoré využívajú klasifikačné modely a sofistikované dataminingové techniky vytvárané na základe dostatočne veľkého objemu dát. Ich veľkou výhodou je to, že na rozdiel od expertných pravidiel je možné pomocou nich nájsť i skryté závislosti v dátach. Sú teda efektívnejšie. Nevýhodou sa potom stáva nutnosť kvalifikovaného nastavenia správneho modelu a jeho vhodná integrácia do dátových systémov poisťovne. Pre efektívne riadenie fraud managementu je teda potrebné skĺbiť dostupné informačné technológie s nástrojmi a metódami na detekciu podvodov. Vhodným metódam prediktívnych modelov sa budeme venovať v nasledujúcich kapitolách tejto práce.

Kapitola 2

Praktická úloha poistenia domácnosti

Na rozdiel od intuitívnej aplikácie expertných pravidiel je praktické použitie prediktívnych modelov mnohokrát komplikované. V prvom rade pre každé poistenie musí byť zvolený jedinečný prediktívny model, ktorý bude odrážať odlišné špecifiká každého poistenia a rôzne údaje, ktoré o klientovi a danom produkte poisťovňa zbiera. Je zrejmé, že iné veličiny budú ovplyvňovať pravdepodobnosť podvodu pri investičnom poistení a iné pri zdravotnom poistení.

Na ukážku prediktívnych modelov sme zo spektra poistných produktov zvolili poistenie domácnosti. Toto poistenie slúži na ochranu vybavenia domácnosti a vecí osobnej potreby. Jedná sa teda o poistenie hnuteľného majetku v trvalo obývanej domácnosti v rodinnom, činžovom, či panelovom dome. Tento typ poistenia chráni proti základným rizikám, ktorými sú živelné katastrofy, poruchy vodovodného zariadenia, či krádeže a vandalizmus. Poisťovne ponúkajú aj poistenie proti špeciálnym rizikám ako napríklad škody spôsobené úderom blesku (prepätie, skrat v elektronický zariadeniach v dôsledku prepätia), rozbitie skla (okná, sklenené plochy, sklo-keramické dosky) a iné [9].

Ďalším problémom aplikácie prediktívnych modelov, na ktorú narážame, je samotný princíp úlohy fraud managementu. Cieľom je definovať model, ktorý umožní na základe historických pozorovaní predikovať pravdepodobnosť podvodu na jednotlivých zmluvách v portfóliu. To ale znamená predpoklad historicky úspešných prípadov odhalenia podvodu. V situácii, keď sa odhaduje, že je podvod poisťovňami správne rozpoznaný len v niekoľkých percentách prípadov, je naplnenie tohto predpokladu sporné. Čiastočným riešením je využitie samoučiacich schopností vybraných prediktívnych modelov, t. j. postupné iterácie k stále úspešnejším modelom.

V našej práci sme k tejto problematike ale pristúpili z odlišného, praktického, pohľadu. Čo poisťovne najviac zaujíma na podvodoch? To, že úspešný podvod znižuje zisky - poistiteľa v konečnom dôsledku zaujíma hlavne ziskovosť jeho portfólia. Z tohto dôvodu spojíme úlohu fraud managementu s úlohou predikcie profitability jednotlivých zmlúv. Vstupom pre našu úlohu sú vybrané reálne údaje o poistení domácnosti.

Náš dátový súbor, viď Príloha A.1, je tvorený 70 žiadosťami o uzatvorenie poistnej zmluvy, rozšírenými o poznatky, ktoré poistiteľ zo žiadosti sám odvodil. Jedná sa napríklad o zaradenie klienta na interný black list alebo o klasifikáciu rizikovosti bydliska. Doplnené sú ďalej údaje o sledovanej profitabilite na zmluve a o tom, či na nej bol detekovaný pokus o podvod. Každý riadok z dátovej matice reprezentuje jednu žiadosť, stĺpce reprezentujú merané znaky. Dáta sú tvorené reálnymi žiadosťami, ktoré boli pre účel tejto práce zúžené na počet postačujúci k praktickej ilustrácii. Pre každú žiadosť máme teda k dispozícii 18 údajov. Jedná sa zväčša o diskrétne, t. j. *kategoriálne* premenné, ktoré nadobúdajú konečný počet hodnôt, napr. `Mesto`. Kategoriálne premenné je možné ďalej rozdeliť na *ordinálne*, ktorých hodnoty sa dajú logicky usporiadať, napr. `Profit`, a *nominálne*, ktoré nemajú logické usporiadanie. Hodnoty priradené kategóriám slúžia len na identifikáciu rôznych kategórií, napr. `Stvrt, Mesto`. Súbor obsahuje tiež *spojité* premenné resp. *kategorizované spojité* premenné, ktoré boli za účelom jednoduchšieho spracovanie prevedené na kategórie, napr. `PC`. Prehľad jednotlivých meraných znakov, ich rozdelenie do kategórii, hodnoty a popis je uvedený v tabuľke 2.1.

Premenná	Kategórie	Hodnoty	Popis
`Profit`	1	PÚ $<$ poistné	Poistná udalosť (PÚ) vs. anualizované poistné po 3 rokoch trvania poistenia.
	2	PÚ $<$ 3 x poistné	
	3	PÚ $<$ 5 x poistné	
	4	PÚ $>$ 5 x poistné	
`Fraud`	1	nie	Odhalený poistný podvod.
	2	áno	
`Stvrt`	1	oblasť s nízkou rizikovosťou	Rozdelenie podľa kvality štvrte. (kriminalita, prírodné živly atď.)
	2	oblasť so zvýšeným rizikom	
	3	riziková oblasť	
`Mesto`	1	Praha	Mesto, oblasť
	2	Kladno	
	3	Neratovice	
	4	Kralupy	
	5	Černošice	
	6	Sázava	
`TO`	1	áno	Trvale obývaná nehnuteľnosť.
	2	nie	
`PC`	1	500	Poistná čiastka v tis. Kč.
	2	750	
	3	1000	
	4	1500	
	5	nad 1500	

`Titul`	1	áno	Vysokoškolský titul
	2	nie	
`Vek`	-	18-99	Vek klienta, celé roky
`Pohlavie`	1	muž	Pohlavie
	2	žena	
`Adr`	1	áno	Adresa rovnaká ako miesto poistenia.
	2	nie	
`SP`	1	áno	Súčasne poistené u iného poistiteľa.
	2	nie	
`Bydlisko`	1	nízko riziková oblasť	Interný skóring adresy.
	2	oblasť so zvýšeným rizikom	
`Platenie`	1	ročne	Frekvencia platenia poistného.
	2	polročne	
	3	štvrťročne	
`BL`	1	áno	Interný Black list.
	2	nie	Zoznam podozrivých klientov.
`DalsieP`	1	nie	Ďalšie poistenie a jeho ziskovosť.
	2	áno - profitabilné	
	3	áno - neprofitabilné	
	4	áno - maximálne neprofitabilné	
`Cudzinec`	1	áno	Cudzinec
	2	nie	
`Zlavy1`	1	áno	Zľavy, vylúčenie vecí mimoriadnej hodnoty.
	2	nie	
`Zlavy2`	1	áno	Zľavy na povodeň, zosuv pôdy, blesky a iné.
	2	nie	

Tabuľka 2.1: Popis dátového súboru

V našom dátovom súbore máme zastúpených 47 mužov a 23 žien. Najmladší poistený je vo veku 23 rokov, najstarší vo veku 72 a celkový vekový priemer je 41,27 rokov. Pokus o poistný podvod bol odhalený v 8 prípadoch.

Z pohľadu profitability (premenná `Profit`) sme zmluvy rozdelili do skupín podľa pomeru poistných plnení z poistných udalostí a objemu násobku ročného poistného (de facto škodný pomer). Za akceptovateľné berieme zmluvy, kde po-

istné plnenie nepresiahlo 3 násobok ročného poistného, t.j. za ziskové považujeme kategórie 1 a 2 premennej `Profit`. Ostatné zmluvy by mali byť z nášho pohľadu z portfólia vylúčené. Podľa nášho kritéria profitability máme v databáze celkom 50 profitabilných poistných zmlúv.

Našou úlohou je teraz nájsť prediktívny model, ktorý bude na nových žiadostiach o poistenie domácnosti identifikovať neziskové zmluvy alebo zmluvy, kde hrozí podvodné jednanie. Výstupom tak bude model, ktorý poisťovni umožní vytriediť nevhodné žiadosti o poistenie už na začiatku procesu poistenia, a teda napomôže k zvýšeniu profitability celého poistného kmeňa.

Kapitola 3

Analýza hlavných komponentov

3.1 Určenie podstatných premenných

Konštrukcia vhodného prediktívneho modelu by mala v prvom kroku zahŕňať dátovú analýzu vstupných premenných a identifikovať tie premenné, ktoré sú podstatné a tie, čo sú pre účel úlohy nadbytočné. Hlavným cieľom v tejto časti práce preto bude vyradiť niektoré premenné z dátovej matice tak, aby sme zachovali vypovedajúcu schopnosť súboru dát. Selekcia podstatných premenných môže totiž mnohokrát viesť k zjednodušeniu výpočtu, väčšej prehľadnosti a jednoduchšej interpretácií navrhnutého modelu a to bez straty dôležitej informácie. Práve týmto problémom sa zaoberá *analýza hlavných komponentov*. Najprv predstavíme metódu teoreticky.

Predpokladajme, že skúmame n objektov, pričom na každom z nich bolo meraných k znakov, značených $X_1, \ldots, X_k$. Hlavným cieľom analýzy je znížiť dimenziu vektoru $\boldsymbol{X} = (X_1, \ldots, X_k)^T$ tak, aby ostala zachovaná čo najväčšia informácia obsiahnutá v celom vektore. Dátová matica má tvar:

$$\begin{pmatrix} x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{nk} \end{pmatrix},$$

kde x_{ij} je hodnotou j-teho znaku na i-tom objekte. Riadky tejto matice predstavujú n pozorovaní vektoru meraných k znakov a označíme ich

$$\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ik})^T, \qquad i = 1, \ldots, n.$$

3.2 Hlavné komponenty

Z vektoru $\boldsymbol{X}$ chceme vytvoriť skrátený vektor $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^T$, $m < k$, ktorého zložky sú vhodnými funkciami zložiek vektoru $\boldsymbol{X}$. Zložky vektoru $\boldsymbol{Z}$ nazývame *hlavné komponenty*. Podľa [7] nové náhodné vektory konštruujeme ako lineárnu kombináciu meraných znakov. Najprv hľadáme vektor $\boldsymbol{C}_1 = (c_{11}, \ldots, c_{1k})^T$ tak, aby náhodná veličina

$$Z_1 = c_{11}X_1 + c_{12}X_2 + \ldots + c_{1k}X_k = \boldsymbol{C}_1^T\boldsymbol{X}$$

mala čo najväčší rozptyl a $\sum_{j=1}^{k} c_{1j}^2 = 1$. To znamená, že Z_1 má medzi všetkými normovanými lineárnymi kombináciami znakov $X_1, \ldots, X_k$ najväčší rozptyl.

Ďalej hľadáme vektor $\boldsymbol{C}_2 = (c_{21}, \ldots, c_{2k})^T$ tak, aby náhodná veličina

$$Z_2 = c_{21}X_1 + c_{22}X_2 + \ldots + c_{2k}X_k = \boldsymbol{C}_2^T \boldsymbol{X}$$

mala čo najväčší rozptyl, bola nekorelovaná so Z_1 a platilo $\sum_{j=1}^{k} c_{2j}^2 = 1$. Takto postupujeme ďalej, až kým zostrojíme všetky hlavné komponenty $Z_1, \ldots, Z_k$. Pre Z_i platí

$$Z_i = c_{i1}X_1 + c_{i2}X_2 + \ldots + c_{ik}X_k = \sum_{j=1}^{k} c_{ij}X_j = \boldsymbol{C}_i^T \boldsymbol{X}, \tag{3.1}$$

kde $\sum_{j=1}^{k} c_{ij}^2 = 1$, Z_i má čo najväčší rozptyl a je nekorelovaná so všetkými už nájdenými Z_j, $j < i$ a $i = 1, \ldots, k$. Týmto postupom teda nájdeme všetky koeficienty matice $\boldsymbol{C}$,

$$\boldsymbol{C} = \begin{pmatrix} c_{11} & \ldots & c_{1k} \\ \vdots & \vdots & \vdots \\ c_{k1} & \ldots & c_{kk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{C}_1^T \\ \vdots \\ \boldsymbol{C}_k^T \end{pmatrix}.$$

Prakticky sa k výpočtu hlavných komponentov využíva spektrálny rozklad kovariančnej matice na vlastné čísla a vlastné vektory. Označme teda kovariančnú maticu vektoru $\boldsymbol{X}$ ako

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1k} \\ \vdots & \sigma_2^2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_k^2 \end{pmatrix},$$

kde $\sigma_i^2 = var(X_i)$ a $\sigma_{ij} = cov(X_i, X_j)$ pre $i \neq j$, kde $i, j = 1, \ldots, k$. Táto matica je symetrická, pretože $cov(X_i, X_j) = cov(X_j, X_i)$.

Podľa [4] existuje pre každú symetrickú maticu spektrálny rozklad v tvare

$$\boldsymbol{\Sigma} = \Gamma \Lambda \Gamma^T = \sum_{j=1}^{k} \lambda_j \gamma_j \gamma_j^T, \tag{3.2}$$

kde

$$\Lambda = diag(\lambda_1, \ldots, \lambda_k)$$

je diagonálna matica a

$$\Gamma = (\gamma_1, \ldots, \gamma_k)$$

je ortogonálna matica zložená z vlastných vektorov $\boldsymbol{\gamma_j}$ príslušných vlastným číslam λ_j matice $\boldsymbol{\Sigma}$. Pre ortogonálnu maticu Γ platí $\Gamma\Gamma^T = I = \Gamma^T\Gamma$, kde I je jednotková matica typu $k \times k$.

Rozptyl i-teho hlavného komponentu môžme za predpokladu $E(\boldsymbol{X}) = 0$ z 3.1 vyjadriť ako

$$var(Z_i) = cov(\boldsymbol{C}_i^T \boldsymbol{X}, \boldsymbol{C}_i^T \boldsymbol{X}) = E(\boldsymbol{C}_i^T (\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{C}_i) = \boldsymbol{C}_i^T \boldsymbol{\Sigma} \boldsymbol{C}_i.$$

Problém maximalizácie rozptylu prvého hlavného komponentu sa redukuje na $\max_{\boldsymbol{C_1}:\|\boldsymbol{C_1}\|=1} \boldsymbol{C}_1^T \boldsymbol{\Sigma} \boldsymbol{C}_1$, kde $\|\boldsymbol{C_1}\|$ je norma vekotru $\boldsymbol{C_1}$. Označme $\boldsymbol{v} = \Gamma^T \boldsymbol{C}_1$. Potom pomocou 3.2 dostaneme

$$\boldsymbol{C}_1^T \Gamma \Lambda \Gamma^T \boldsymbol{C}_1 = \boldsymbol{v}^T \Lambda \boldsymbol{v} = \sum_{j=1}^{k} v_j^2 \lambda_j.$$

Vzhľadom k tomu, že Γ je ortogonálna matica, platí $\|\boldsymbol{C_1}\| = \|\boldsymbol{v}\|$. Ďalej položme $m_j = v_j^2$ a problém maximalizácie sa zmení na

$$\max_{m_j \geq 0, \sum_{j=1}^{k} m_j = 1} \sum_{j=1}^{k} m_j \lambda_j. \tag{3.3}$$

Je zjavné, že

$$\max_{m_j \geq 0, \sum_{j=1}^{k} m_j = 1} \sum_{j=1}^{k} m_j \lambda_j \leq \lambda_{j^*} \max_{m_j \geq 0, \sum_{j=1}^{k} m_j = 1} \sum_{j=1}^{k} m_j = \lambda_{j^*},$$

kde indexom j^* označíme maximálne vlastné číslo matice $\boldsymbol{\Sigma}$. Usporiadame vlastné čísla zostupne, t. j. $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k$. Maximum 3.3 získame, ak m_1 odpovedajúce najväčšiemu vlastnému číslu λ_1 je 1 a $m_j = 0$ pre $j = 2, \ldots, k$. Potom spätne $\boldsymbol{C}_1$ je rovné vlastnému vektoru γ_{j^*} matice Γ, príslušnému najväčšiemu vlastnému číslu, teda λ_1. Vektor $\boldsymbol{C_1}$ je ortonormálny a maximálny rozptyl pre prvý hlavný komponent získame pre $\boldsymbol{C}_1 = \gamma_1$. Hodnota maxima je λ_1. Keďže vlastné vektory symetrickej matice sú ortogonálne, k nájdeniu druhého hlavného komponentu maximalizujeme rozptyl tak, že $v_1 = 0$. Hľadáme $\boldsymbol{C}_2$ ortogonálny k $\boldsymbol{C}_1$. Ide o rovnaký maximalizačný problém ako 3.3 s dodatočnou podmienkou $m_1 = 0$. Maximum podla predošlého získame, keď vektor $\boldsymbol{C}_2$ bude rovný vlastnému vektoru prislúchajúceho druhému najväčšiemu vlastnému číslu matice $\boldsymbol{\Sigma}$, v našom usporiadaní λ_2. Analogicky postupujeme ďalej a pre i-ty hlavný komponent dostávame, že $\boldsymbol{C}_i$ sú vlastné vektory γ_i matice $\boldsymbol{\Sigma}$ príslušné vlastným číslam λ_i a $var(Z_i) = \lambda_i$, $i = 1, \ldots, k$.

Koeficienty $c_{i1}, \ldots, c_{ik}$ vektoru $\boldsymbol{C}_i$ potom hľadáme ako normované riešenie sústavy rovníc

$$(\boldsymbol{\Sigma} - \lambda_i \boldsymbol{I})\boldsymbol{C}_i = 0$$

tak, že λ_i je i-ty najväčší koreň rovnice

$$det(\boldsymbol{\Sigma} - \lambda \boldsymbol{I}) = 0. \tag{3.4}$$

Kovariančná matica hlavných komponentov $\boldsymbol{\Sigma_Z}$ teda bude tvaru

$$\boldsymbol{\Sigma_Z} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{pmatrix}.$$

Keďže transformačná matica $\boldsymbol{C}$ je ortogonálna, vieme pomocou získaných hlavných komponentov zo vzťahu $\boldsymbol{Z} = \boldsymbol{CX}$ vyjadriť pôvodné náhodné veličiny X_i ako

$$X_i = \sum_{j=1}^{k} c_{ij} Z_j.$$

Pre hlavné komponenty platí, že celková variabilita meraných znakov je popísaná celkovou variabilitou hlavných komponentov. Pritom hlavné komponenty s nižšími indexmi vysvetľujú väčšiu časť z celkovej variability a prvý hlavný komponent popisuje najväčšiu časť rozptylu pôvodných premenných

$$\sum_{j=1}^{k} var(X_j) = \sum_{j=1}^{k} var(Z_j) = \sum_{j=1}^{k} \lambda_j.$$

3.2.1 Počet hlavných komponentov a štandardizácia dát

V úlohe o znížení dimenzie vektoru charakteristických znakov je dôležitá aj voľba počtu hlavných komponentov. Chceme dosiahnuť, aby počet hlavných komponentov bol menší ako počet pozorovaných znakov, ale zároveň tak veľký, že celková variabilita hlavných komponentov odráža dostatočne veľkú časť variability meraných znakov. K voľbe počtu hlavných komponentov sa podľa [7] využíva pomer variability hlavných komponentov a celkovej premenlivosti meraných znakov

$$q(m) = \frac{\sum_{j=1}^{m} var(Z_j)}{\sum_{j=1}^{k} var(X_j)}.$$

Z priebehu krivky $q(m)$ volíme počet hlavných komponentov tak, aby

$$q(m) > 1 - \ \epsilon$$

alebo

$$q(m+1) - q(m) < \epsilon$$

pre dané ϵ. Zväčša stačí k popisu variability 2 až 5 hlavných komponentov. Pokiaľ $q(m)$ rastie pomaly, naznačuje to nevhodnú voľbu znakov a analýzu je treba opakovať pridaním nových znakov alebo s úplne novým súborom znakov.

Problémom pri praktických výpočtoch býva fakt, že vo väčšine prípadov nie je známa kovariančná matica $\boldsymbol{\Sigma}$ základného súboru znakov. V prípadoch, keď matica $\boldsymbol{\Sigma}$ nie je známa, sa používa výberová kovariančná matica $\widehat{\boldsymbol{\Sigma}}$, ktorej prvky sú tvaru

$$\hat{\sigma}_{ij} = \frac{1}{n-1} \sum_{t=1}^{n} (x_{ti} - \overline{x}_i)(x_{tj} - \overline{x}_j),$$

kde $\overline{x}_i = \frac{1}{n}\sum_{s=1}^{n} x_{si}$ je priemer i-tej premennej cez všetky objekty. Hlavné komponenty, ktoré získame pomocou výberovej kovariančnej matice sa nazývajú *výberové hlavné komponenty*, značíme ich $\hat{Z}_1, \ldots, \hat{Z}_k$.

Pokiaľ sú pozorované znaky merané v rôznych jednotkách, využívajú sa tzv. *štandardizované premenné*, ktoré už nie sú závislé na jednotkách pôvodných premenných. Získame ich transformáciou

$$X_j^* = \frac{X_j - EX_j}{\sqrt{varX_j}}, \qquad j = 1, \ldots, k. \tag{3.5}$$

Takto transformované premenné majú nulovú strednú hodnotu a jednotkový rozptyl. Požadujeme, aby platilo

$$k = \sum_{j=1}^{k} var(X_j^*) = \sum_{j=1}^{k} var(Z_j). \tag{3.6}$$

Ďalej pre určenie počtu hlavných komponentov v prípade štandardizovaných premenných platí

$$q(m) = \frac{\sum_{j=1}^{m} \lambda_j}{k}. \tag{3.7}$$

Nech $\mathbf{\Sigma}^*$ je kovariančná matica štandardizovaných premenných. Potom pre túto maticu platí, že jej diagonálne prvky sú rovné jednej a je súčasne korelačnou maticou pôvodných pozorovaných znakov.

Pri vyhodnocovaní výsledkov analýzy hlavných komponentov sa väčšia pozornosť kladie na veľkosť absolútnych hodnôt koeficientov $c_{i1}, c_{i2}, \ldots, c_{ik}$ jednotlivých komponentov. Dôvodom je najmä náročná interpretácia samotných hlavných komponentov. Získané koeficienty $c_{i1}, c_{i2}, \ldots, c_{ik}$ nadobúdajú hodnoty v intervale $\langle -1, 1\rangle$ a predstavujú váhy znakov $X_1, \ldots, X_k$ v i-tom hlavnom komponente. Z dôvodu, že prvý hlavný komponent vysvetľuje najväčšiu časť celkovej variability, sú jeho znaky s veľkou absolútnou hodnotou váh tie najdôležitejšie k popisu skúmaných objektov. Podobne sú podľa absolútnej hodnoty koeficientov vyberané dôležité znaky v druhom hlavnom komponente atď. Popísaným postupom nám teda analýza hlavných komponentov umožňuje rozlišovať podstatné merané znaky pre skúmané objekty.

3.3 Aplikácia na dátovú maticu

Analýzu hlavných komponentov teraz využijeme v našej úlohe tvorby prediktívneho modelu riadenia profitability a podvodov, na základe žiadostí o poistenie domácnosti tak, ako sme ich definovali v predchádzajúcej kapitole. Účelom bude znížiť rozmer úlohy a tým vylepšiť jej použiteľnosť v reálnych situáciách. K spracovaniu dát využijeme softvér *Wolfram Mathematica 8.0*, [15], v ktorom sme vytvorili vlastný kód pre spracovanie úlohy. Kód je písaný všeobecným spôsobom s použitím parametrizácie na vstupnom súbore, aby sme ilustrovali možnosť jednoduchého prispôsobenia v prípade aplikácie na iný typ poistenia a iný dátový súbor. Okomentovaný kód z *Mathematici* prikladáme v Prílohe B.1.

Na začiatku analýzy je vhodné spočítať výberové priemery a smerodajné odchýlky pozorovaných znakov. Hodnoty sú uvedené v tabuľke 3.1.

Premenná	Výberový priemer	Výberová smerodajná odchýlka
Profit	1,92857	1,08108
Fraud	1,11429	0,320455
Stvrt	2,07143	0,767481
Mesto	3,25714	1,76673
TO	1,17143	0,379604
PC	3,1	1,20566
Titul	1,61429	0,490278
Vek	41,2714	11,8553
Pohlavie	1,32857	0,473085
Adr	1,38571	0,490278
SP	1,8	0,402888
Bydlisko	1,05714	0,233791
Platenie	1,47143	0,696189
BL	1,82857	0,379604
DalsieP	1,9	0,98024
Cudzinec	1,92857	0,259399
Zlavy1	1,51429	0,503405
Zlavy2	1,37143	0,486675

Tabuľka 3.1: Výberový priemer a smerodajná odchýlka

Vidíme, že priemery a smerodajné odchýlky jednotlivých premenných sú rádovo odlišné, je preto vhodné dáta štandardizovať ako sme uviedli v časti 3.2.1. Využijeme transformáciu dát podľa 3.5. Vo všetkých našich výpočtoch zachovávame poradie premenných, ktoré je uvedené v tabuľke 2.1.

Ďalej k výpočtu hlavných komponentov potrebujeme určiť kovariančnú maticu štandardizovaných premenných, ktorá je zároveň korelačnou maticou pôvodných meraných znakov. Maticu uvádzame v tabuľke 3.2 s hodnotami zaokrúhlenými na dve desatinné miesta. Vysokú korelovanosť môžme pozorovať medzi premennými Stvrt a PC ($|0,71|$), kde vyššia rizikovosť štvrte môže prispievať k poisteniu na vyššie poistné sumy. Keďže sa zväčša od poistnej čiastky odvíja výška poistného, frekvenčné zľavy, zľavy za objem poistného, rôzne provízie a v konečnom dôsledku teda aj profit poisťovne, je vysoká korelovanosť medzi premennými PC a Profit vcelku očakávaná. Ďalšími významne korelovanými znakmi sú TO a ADR, Zlavy1 a ADR, či DalsieP a BL.

V ďalšom kroku analýzy spočítame vlastné čísla kovariančnej matice, ktoré vyjadrujú rozptyl príslušných hlavných komponentov. Podľa pomeru celkovej variability a variability, ktorú popisujú hlavné komponenty, stanovíme počet hlavných komponentov. Vychádzame pritom z rovníc 3.4 a 3.7.

V tabuľke 3.3 je uvedená hodnota vlastných čísel spočítaných softvérom. Taktiež je v nej vyjadrené koľko percent tvorí dané vlastné číslo v ich celkovom súčte.

	Profit	Fraud	Stvrt	Mesto	TO	PC	Titul	Vek	Pohlavie	Adr	SP	Bydlisko	Platenie	BL	DalsieP	Cudzinec	Zlavy1	Zlavy2
Profit	1,	0,19	0,43	0,37	0,14	-0,6	0,19	-0,08	-0,12	0,27	-0,07	0,07	0,32	-0,24	0,08	0,03	-0,33	-0,36
Fraud	0,19	1,	0,32	0,	0,07	-0,18	0,1	-0,15	-0,06	0,18	0,07	0,1	0,21	-0,31	-0,01	-0,07	-0,01	-0,18
Stvrt	0,43	0,32	1,	0,11	0,16	-0,71	0,27	-0,2	-0,11	0,27	0,	-0,02	0,26	-0,16	0,05	0,1	-0,25	-0,27
Mesto	0,37	0,	0,11	1,	-0,15	-0,39	0,17	-0,03	-0,02	-0,1	0,01	0,07	0,05	-0,04	0,15	0,04	-0,12	-0,05
TO	0,14	0,07	0,16	-0,15	1,	-0,16	-0,03	-0,08	-0,08	0,57	0,04	0,21	-0,09	0,01	-0,11	0,13	-0,24	-0,19
PC	-0,6	-0,18	-0,71	-0,39	-0,16	1,	-0,25	0,19	0,04	-0,26	-0,08	-0,12	-0,23	0,23	-0,16	-0,02	0,3	0,28
Titul	0,19	0,1	0,27	0,17	-0,03	-0,25	1,	0,02	0,05	0,09	0,04	0,07	0,12	-0,05	0,07	-0,11	-0,07	-0,06
Vek	-0,08	-0,15	-0,2	-0,03	-0,08	0,19	0,02	1,	0,15	-0,05	0,08	-0,06	0,26	0,37	-0,21	0,12	0,31	0,14
Pohlavie	-0,12	-0,06	-0,11	-0,02	-0,08	0,04	0,05	0,15	1,	-0,24	0,2	-0,04	-0,04	0,	-0,02	-0,04	0,13	0,09
Adr	0,27	0,18	0,27	-0,1	0,57	-0,26	0,09	-0,05	-0,24	1,	-0,04	0,18	0,14	-0,03	-0,04	0,11	-0,52	-0,37
SP	-0,07	0,07	0,	0,01	0,04	-0,08	0,04	0,08	0,2	-0,04	1,	-0,03	0,03	0,15	0,02	0,	0,01	0,01
Bydlisko	0,07	0,1	-0,02	0,07	0,21	-0,12	0,07	-0,06	-0,04	0,18	-0,03	1,	0,1	0,11	-0,1	0,07	-0,13	-0,19
Platenie	0,32	0,21	0,26	0,05	-0,09	-0,23	0,12	0,26	-0,04	0,14	0,03	0,1	1,	0,15	0,03	0,11	-0,08	-0,14
BL	-0,24	-0,31	-0,16	-0,04	0,01	0,23	-0,05	0,37	0,	-0,03	0,15	0,11	0,15	1,	-0,47	0,02	0,01	0,11
DalsieP	0,08	-0,01	0,05	0,15	-0,11	-0,16	0,07	-0,21	-0,02	-0,04	0,02	-0,1	0,03	-0,47	1,	-0,03	0,02	0,11
Cudzinec	0,03	-0,07	0,1	0,04	0,13	-0,02	-0,11	0,12	-0,04	0,11	0,	0,07	0,11	0,02	-0,03	1,	-0,05	-0,13
Zlavy1	-0,33	-0,01	-0,25	-0,12	-0,24	0,3	-0,07	0,31	0,13	-0,52	0,01	-0,13	-0,08	0,01	0,02	-0,05	1,	0,33
Zlavy2	-0,36	-0,18	-0,27	-0,05	-0,19	0,28	-0,06	0,14	0,09	-0,37	0,01	-0,19	-0,14	0,11	0,11	-0,13	0,33	1.

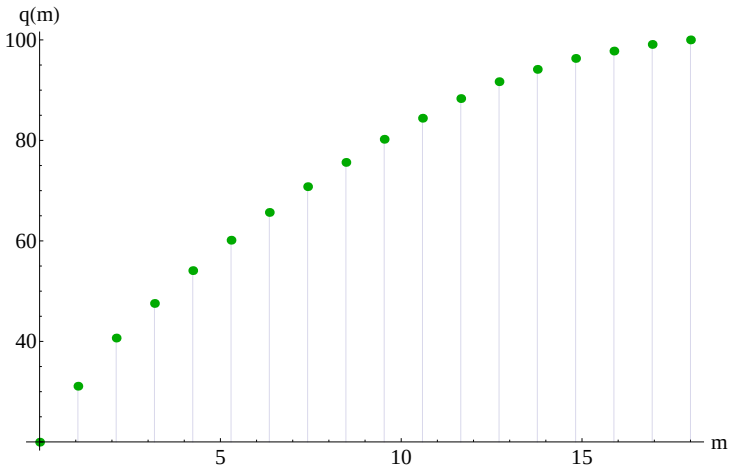
Tabuľka 3.2: Kovariančná matica štandardizovaných premenných

Uvádzame v nej tiež kumulatívne súčty vlastných čísel v celkovej variabilite.

Vlastné číslo	Hodnota vl. čísla	Percentuálna časť	Kumulatívne percentá
1	3,59036	19,9464	19,9464
2	2,00474	11,1374	31,0839
3	1,72534	9,58522	40,6691
4	1,23934	6,8852	47,5543
5	1,17605	6,53362	54,0879
6	1,089	6,05	60,1379
7	0,995178	5,52876	65,6667
8	0,92267	5,12594	70,7926
9	0,869463	4,83035	75,6229
10	0,828219	4,60122	80,2242
11	0,754053	4,18918	84,4133
12	0,70441	3,91339	88,3267
13	0,604434	3,35797	91,6847
14	0,441436	2,45242	94,1371
15	0,392073	2,17819	96,3153
16	0,263601	1,46445	97,7798
17	0,237636	1,3202	99,1
18	0,162007	0,90004	100

Tabuľka 3.3: Vlastné čísla kovariančnej matice

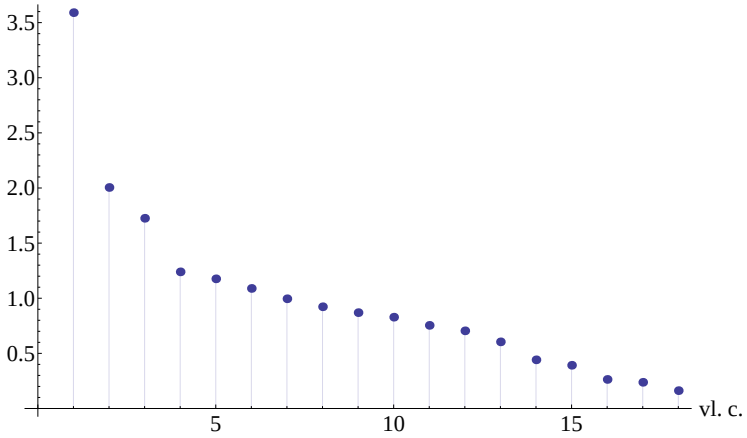
Prvý hlavný komponent vysvetľuje takmer 19,95%, prvých 5 komponentov už viac než 54 % celkovej variability. Z grafu funkcie $q(m)$, viď obrázok 3.1, vidíme, že funkcia rastie dostatočne rýchlo pre prvých 5 hlavných komponentov. Práve týchto 5 hlavných komponentov budeme považovať za významné pre popis celkovej variability.



Obr. 3.1: Funckia q(m)

Aby sme si potvrdili, že nami zvolený počet je dostatočný, uvádzame taktiež obrázok 3.2 zostupne zoradených vlastných čísiel, tzv. *scree graf*. Tento graf od-

borná literatúra tiež odporúča na určenie počtu komponentov. Vidíme, že hodnota vlastných čísel klesá rýchlejšie najmä pre prvých 5 komponentov, a teda k ďalšej analýze využijeme len 5 prvých komponentov.



Obr. 3.2: Scree graf

K vypočítaným vlastným číslam následne určíme normované vlastné vektory a získame tak maticu koeficientov hlavných komponentov. Podstatnými koeficientami pre nás budú tie, ktoré v prvých 5 hlavných komponentoch nadobúdajú v absolútnej hodnote vyššiu hodnotu ako 0,25. Koeficienty sú uvedené v tabuľke 3.4. Overíme tiež, že celková variabilita meraných znakov je popísaná celkovou variabilitou hlavných komponentov. Najprv teda spočítame rozptyly hlavných komponentov: 3,59036; 2,00474; 1,72534; 1,23934; 1,17605; 1,089; 0,995178; 0,92267; 0,869463; 0,828219; 0,754053; 0,70441; 0,604434; 0,441436; 0,392073; 0,263601; 0,237636; 0,162007. Ich hodnoty odpovedajú vlastným číslam kovariančnej matice. Podľa rovnice 3.6 určíme ich celkový súčet, ktorý je podľa očakávania rovný súčtu hodnôt vlastných čísiel a počtu meraných znakov, teda 18.

Najdôležitejšími premennými pre celkovú variabilitu dát z prvého hlavného komponentu sú `Profit`, `Stvrt`, `PC`, `ADR`, `Zlavy1` a `Zlavy2`. Premennými, ktoré sa javia nevýznamné z pohľadu nami vybraných komponentov, sú `Cudzinec`, `Bydlisko` a `Titul`. U premennej `Titul` jej nepodstatný význam môžme interpretovať tak, že ľudia mnohokrát vysokoškolský titul do žiadostí neuvádzajú a teda majú automaticky vyplnenú možnosť, že titul nemajú i keď ho v skutočnosti získali. Uzavrieť poistenie je pre cudzincov zdĺhavejší proces a pokiaľ sú v danej krajine krátkodobo poistenie zväčša neuzatvárajú, preto sa premenná `Cudzinec` môže javiť nevýznamnou. Premenná `Bydlisko` vyjadruje interný skóring rizikovosti bydliska poistenia domácnosti. Jej nevýznamnosť z titulu variability nutne neznamená, že je nevýznamná vo všeobecnosti, ale len v prípade nami použitej dátovej matice. Je nutné si uvedomiť, že z praktických dôvodov používame malú vzorku vybraných dát. Pre ďalšie spracovanie úlohy tak na základe výstupu analýzy hlavných komponentov môžme vynechať 3 merané znaky `Titul`, `Cudzinec` a `Bydlisko`, čím dosiahneme zníženie rozmeru úlohy na 15 z pohľadu variability podstatných znakov.

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{15}	Z_{16}	Z_{17}	Z_{18}
Profit	0,385	-0,107	-0,165	0,177	-0,043	0,012	0,005	0,077	0,065	0,181	0,33	0,115	-0,078	0,614	-0,074	-0,383	-0,268	0,087
Fraud	0,208	-0,093	0,027	-0,524	-0,344	0,187	-0,26	0,146	-0,125	-0,134	0,11	0,191	0,442	-0,154	0,016	-0,281	0,016	-0,222
Stvrt	0,378	-0,1	-0,127	-0,148	-0,094	-0,038	0,263	0,157	0,173	-0,348	-0,14	-0,321	-0,103	-0,233	-0,085	-0,167	0,064	0,572
Mesto	0,157	-0,267	-0,256	0,439	0,276	0,034	-0,219	0,049	-0,081	-0,142	0,272	0,276	0,311	-0,31	0,264	0,176	0,031	0,218
TO	0,198	0,382	0,228	-0,183	0,169	-0,194	0,051	-0,276	0,146	-0,192	0,389	0,000	-0,037	0,129	0,455	-0,044	0,384	-0,002
PC	-0,412	0,174	0,17	-0,027	-0,153	0,064	-0,075	-0,053	-0,066	0,211	-0,144	0,246	0,218	0,122	0,207	-0,229	-0,015	0,668
Titul	0,158	-0,148	-0,243	-0,148	0,194	0,31	0,148	-0,476	0,297	-0,059	-0,49	0,333	0,067	0,16	0,106	0,016	0,032	-0,072
Vek	-0,171	0,193	-0,487	-0,016	-0,195	-0,146	0,085	-0,257	0,13	0,136	0,298	0,249	-0,06	-0,235	-0,455	-0,122	0,309	0,037
Pohlavie	-0,122	-0,107	-0,203	-0,346	0,383	-0,221	-0,205	0,078	0,436	0,421	0,069	-0,325	0,21	-0,09	0,094	-0,051	-0,164	0,025
Adr	0,318	0,37	0,164	-0,071	-0,015	-0,066	0,193	-0,271	-0,061	0,15	0,092	0,118	0,156	-0,263	-0,132	0,261	-0,615	0,095
SP	-0,017	0,011	-0,216	-0,426	0,434	-0,299	0,029	0,16	-0,536	-0,124	-0,072	0,257	-0,138	0,184	-0,118	0,101	-0,016	0,127
Bydlisko	0,124	0,241	-0,027	0,031	0,142	0,255	-0,725	-0,3	-0,118	-0,123	-0,066	-0,269	-0,155	0,039	-0,264	-0,032	0,004	0,151
Platenie	0,167	0,053	-0,451	-0,053	-0,414	-0,019	-0,005	-0,103	-0,311	0,297	-0,086	-0,306	0,042	0,144	0,369	0,348	0,115	0,038
BL	-0,167	0,411	-0,364	0,145	0,169	0,185	0,172	0,084	-0,174	-0,152	-0,077	-0,151	-0,014	-0,185	0,287	-0,491	-0,249	-0,219
DalsieP	0,07	-0,428	0,173	0,049	-0,017	-0,338	-0,015	-0,48	-0,32	0,181	-0,067	-0,077	-0,168	-0,263	0,114	-0,419	-0,026	-0,051
Cudzinec	0,059	0,184	-0,088	0,223	-0,193	-0,668	-0,247	0,036	0,181	-0,274	-0,382	0,088	0,241	0,133	-0,011	-0,049	-0,082	-0,115
Zlavy1	-0,305	-0,19	-0,151	-0,201	-0,27	-0,034	-0,177	-0,112	0,194	-0,354	0,227	0,12	-0,434	-0,013	0,262	0,145	-0,426	0,043
Zlavy2	-0,297	-0,179	-0,037	0,005	0,038	0,001	0,225	-0,343	-0,133	-0,36	0,216	-0,367	0,496	0,286	-0,199	0,049	-0,092	0,024

Tabuľka 3.4: Matica koeficientov hlavných komponentov

Kapitola 4

Diskriminačná analýza

4.1 Ciele diskriminačnej analýzy

Úlohou fraud managementu je predikovať pravdepodobnosť pokusu o podvod. V kontexte našej úlohy sa zameriavame na nové žiadosti o poistenie domácnosti, na základe ktorých je našim cieľom odhadnúť budúcu profitabilitu so zahrnutím pokusov o podvod medzi neprofitabilné zmluvy. Pomocou analýzy hlavných komponentov sme z nášho dátového súboru vybrali premenné, ktoré nie sú významné z pohľadu zachovania celkovej variability dátového súboru. Pokiaľ sa naša databáza rozširuje príchodom nových žiadateľov, je našim zámerom zaradiť novú zmluvu do niektorej z už existujúcich skupín podľa úrovne profitability a poistného podvodu. A to na základe sledovaných kvantitatívnych premenných. Práve pre tento účel využijeme metódu *diskriminačnej analýzy*.

Podľa [5] je prvotnou úlohou tejto analýzy skúmať schopnosť sledovaných premenných separovať jednotlivé skupiny v súbore. V našom prípade profitabilné zmluvy od tých neporofitabilných. Táto metóda teda analyzuje závislosti medzi skupinou nezávisle premenných (diskriminátorov) a jednou závisle premennou. Budeme postupovať obdobne ako v predchádzajúcej kapitole. Diskriminačnú analýzu najprv popíšeme teoreticky, pričom budeme vychádzať najmä z [5] a na klasifikáciu pozorovaní využijeme [1]. Následne budeme aplikovať túto analýzu na náš dátový súbor s významnými premennými, ktoré sme získali pomocou analýzy hlavných komponentov.

4.2 Kanonická diskriminačná analýza

Opäť využijeme predpoklad, že skúmame n objektov pomocou k pozorovaných znakov. V tejto fáze analýzy však už máme objekty rozdelené do M skupín značených $K_1, \ldots, K_M$. Budeme teda pracovať s dátovou maticou $\boldsymbol{X}$, ktorej riadky sú tvorené vektormi $\boldsymbol{X}_{mi} = (X_{mi1}, \ldots, X_{mik})^T$, kde $m = 1, \ldots, M$ vyjadruje príslušnosť k skupine a $i = 1, \ldots, n_m$ počet objektov v danej skupine. Teda n_m vyjadruje počet objektov v skupine K_m.

V deskriptívnej časti diskriminačnej analýzy hľadáme lineárnu kombináciu k pozorovaných znakov

$$Y = \boldsymbol{c}^T \boldsymbol{X},$$

tak, aby najlepšie odlišovala M skupín. A to v zmysle, že jej vnútroskupinová variabilita je čo najmenšia a medziskupinová variabilita čo najväčšia. Vnútroskupinovú variabilitu označíme

$$\boldsymbol{E} = \sum_{m=1}^{M} \sum_{i=1}^{n_m} (\boldsymbol{X}_{mi} - \overline{\boldsymbol{X}}_m)(\boldsymbol{X}_{mi} - \overline{\boldsymbol{X}}_m)^T = (n - M)\boldsymbol{S}. \tag{4.1}$$

A medziskuponovú variabilitu ako maticu tvaru

$$\boldsymbol{B} = \sum_{m=1}^{M} \sum_{i=1}^{n_m} (\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{X}})^T = \sum_{m=1}^{M} n_m (\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{X}})^T. \tag{4.2}$$

Celkovú variabilitu potom vyjadruje súčet matíc $\boldsymbol{E} + \boldsymbol{B} = \boldsymbol{T}$,

$$\boldsymbol{T} = \sum_{m=1}^{M} \sum_{i=1}^{n_m} (\boldsymbol{X}_{mi} - \overline{\boldsymbol{X}})(\boldsymbol{X}_{mi} - \overline{\boldsymbol{X}})^T,$$

kde vektor priemerov a výberová kovariančná matica v m-tej skupine sú

$$\overline{\boldsymbol{X}}_m = \frac{1}{n_m} \sum_{i=1}^{n_m} \boldsymbol{X}_{mi}, \qquad \boldsymbol{S}_m = \frac{1}{n_m - 1} \sum_{i=1}^{n_m} (\boldsymbol{X}_{mi} - \overline{\boldsymbol{X}}_m)(\boldsymbol{X}_{mi} - \overline{\boldsymbol{X}}_m)^T.$$

Celkový priemer a výberová kovariančná matica sú tvaru

$$\overline{\boldsymbol{X}} = \frac{1}{n} \sum_{m=1}^{M} \overline{\boldsymbol{X}}_m n_m, \qquad \boldsymbol{S} = \frac{1}{n - M} \sum_{m=1}^{M} \boldsymbol{S}_m (n_m - 1).$$

Variabilitu Y teda môžme vyjadriť v tvare $\boldsymbol{c}^T\boldsymbol{T}\boldsymbol{c} = \boldsymbol{c}^T\boldsymbol{E}\boldsymbol{c} + \boldsymbol{c}^T\boldsymbol{B}\boldsymbol{c}$. Maximum medziskupinovej a minimum vnútroskupinvej varibility Y získame maximalizáciou tzv. *Fisherovho diskriminačného kritéria*

$$\lambda = \frac{\boldsymbol{c}^T\boldsymbol{B}\boldsymbol{c}}{\boldsymbol{c}^T\boldsymbol{E}\boldsymbol{c}}. \tag{4.3}$$

Hľadáme teda prvky vektoru $\boldsymbol{c}$. A to tak, že položíme parciálne derivácie λ podľa $\boldsymbol{c}$ rovné 0. Získame sústavu rovníc tvaru

$$(\boldsymbol{B} - \lambda\boldsymbol{E})\boldsymbol{c} = 0 \tag{4.4}$$

$$(\boldsymbol{B}\boldsymbol{E}^{-1} - \lambda\boldsymbol{I})\boldsymbol{c} = 0, \qquad |\boldsymbol{E}| \neq 0.$$

Sústava má l nenulových riešení, pokiaľ je charakteristický polynóm $\left|\boldsymbol{B}\boldsymbol{E}^{-1} - \lambda\boldsymbol{I}\right|$ rovný nule. Jeho riešením sú vlastné čísla $\lambda_1, \ldots, \lambda_l$ matice $\boldsymbol{B}\boldsymbol{E}^{-1}$. Počet riešení je $l = \min(k, M-1)$.Vlastným číslam odpovedajú ortogonálne vlastné vektory. Vektor $\boldsymbol{c}_1$ odpovedá najväčšiemu vlastnému číslu, pokiaľ vlastné čísla usporiadame zostupne $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_l$. Jednoznačné riešenie $\boldsymbol{c}_1$ možno získať normovaním

$$\boldsymbol{c}_1^T\boldsymbol{E}\boldsymbol{c}_1 = 1 \tag{4.5}$$

alebo

$$\frac{1}{(n-M)}\boldsymbol{c}_1^T\boldsymbol{E}\boldsymbol{c}_1 = 1. \tag{4.6}$$

Maximum 4.3 v prípade 4.5 vieme vyjadriť ako

$$\lambda_1 = \frac{\boldsymbol{c}_1^B\boldsymbol{B}\boldsymbol{c}_1}{1},$$

resp. z 4.6 ako

$$\lambda_1 = \frac{1}{(n-M)}\frac{\boldsymbol{c}_1^B\boldsymbol{B}\boldsymbol{c}_1}{1}.$$

$Y_1 = \boldsymbol{c}_1^T\boldsymbol{X}$ sa nazýva *prvý diskriminant* alebo *prvá kanonická premenná*. Jej medziskupinovú variabilitu vyjadruje λ_1. Ďalšie kanonické premenné získame na základe zvyšných vlastných čísiel a im príslušných vlastných vektorov, teda $Y_i = \boldsymbol{c}_i^T\boldsymbol{X}, i = 2, 3, \ldots, l$. Prvky vektoru $\boldsymbol{c}_i$ sú potom koeficientami i-tej kanonickej premennej a dosadením pozorovaných hodnôt za $X_1, \ldots, X_k$ získame tzv. i-te *diskriminačné skóre*.

Pričítaním konštanty $v_i = -\boldsymbol{c}_i^T\overline{\boldsymbol{X}}$ k i-tému diskriminačnému skóre pozorovania charakterizovaného vektorom $\boldsymbol{X}$, t. j. vytvorením atribútu

$$y_i = v_i + \boldsymbol{c}_i^T\boldsymbol{X},$$

dosiahneme, že priemerné diskriminačné skóre jednotlivých diskriminantov bude nulové.

K analýze toho ako sa vzhľadom k i-tej kanonickej premennej jednotlivé skupiny vzájomne líšia sa využíva vektor priemerných hodnôt diskriminantov v skupinách

$$\overline{y}_{mi} = v_i + \sum_{j=1}^{k} c_{ij}\overline{X}_{mj}, \qquad i = 1, \ldots, l. \tag{4.7}$$

Koeficient c_{ij} vyjadruje vplyv j-tej pôvodnej premennej na i-tú kanonickú premennú Y_i. Tieto koeficienty bývajú normované na tvar

$$\boldsymbol{c}_i^* = \frac{1}{\sqrt{n-M}}\boldsymbol{H}\boldsymbol{c}_i, \tag{4.8}$$

kde $\boldsymbol{H}$ je diagonálna matica tvorená odmocninami diagonálnych prvkov matice $\boldsymbol{E}$.

Ďalšou charakteristikou, ktorá sa využíva k posúdeniu významnosti premennej pre daný diskriminant, je korelačný koeficient medzi pôvodnými premennými a kanonickou premennou

$$\boldsymbol{r}_i = \frac{1}{\sqrt{n-M}}\boldsymbol{H}^{-1}\boldsymbol{E}\boldsymbol{c}_i. \tag{4.9}$$

Vysoká absolútna hodnota tohto koeficientu naznačuje, že premenná je pre diskriminant významná.

K overeniu hypotézy, že aspoň jedna kanonická premenná je dôležitá k odlíšeniu jednotlivých skupín sa využíva tzv. *Wilksova štatistika* $\Lambda = |\boldsymbol{E}| \,/\, |\boldsymbol{E} + \boldsymbol{B}|$

pre $l < 2$. Test hypotézy o nulovosti vlastných čísiel $\lambda_1 = \lambda_2 = \ldots, \lambda_l = 0$ je ekvivalentný testu na zhodu vektorov stredných hodnôt M skupinových vektorov. Pre $l \geq 2$ sa používa tzv.*Bartlettova štatistika*

$$V = c(-\ln \Lambda), \qquad c = n - 1 - (k + M)/2,$$

ktorá má približne chí-kvadrát rozdelenie o $k(M-1)$ stupňoch voľnosti. Predpokladáme pritom nezávislosť skupín, normalitu skupín a zhodu kovariančných matíc.

4.3 Klasifikácia pomocou diskriminačnej analýzy

Druhou hlavnou úlohou diskriminačnej analýzy je klasifikácia objektov do už existujúcich skupín. Predpokladajme podľa [5], že sme zvolili q prvých diskriminantov. Na zaradenie i-teho pozorovania do skupiny sa využíva vzdialenosť objektu od skupinového centroidu. Vzdialenosť s-tého objektu od m-tej skupiny možno teda vyjadriť ako

$$d^2_{sm} = \sum_{i=1}^{q} [\boldsymbol{v}_i^T (\boldsymbol{X}_s - \overline{\boldsymbol{X}}_m)]^2 = \sum_{i=1}^{q} (y_{si} - \overline{y}_{mi}), \tag{4.10}$$

kde y_{si} je i-té diskriminačné skóre s-tého pozorovania v súbore a y_{mi} je priemer určený podľa 4.7. Pozorovanie zaradíme do tej skupiny, pre ktorú je vzdialenosť d^2_{sm} najmenšia. Nevýhodou tejto klasifikácie je podľa odbornej literatúry najmä to, že sa neberie do úvahy veľkosť jednotlivých skupín. A teda sa opomína fakt, že apriórne pravdepodobnosti príslušnosti k skupinám sa môžu líšiť.

Existuje alternatívny prístup roztriedenia pozorovaní do existujúcich skupín. A to na základe rozhodovacieho pravidla, resp. funkcie, ktorá sa odhaduje z pôvodných premenných a objektov už zaradených do skupín. Predpokladajme, že poznáme vektor meraných znakov nového pozorovania $\boldsymbol{X} = (X_1, \ldots, X_k)^T$, kde $\boldsymbol{X} \in \boldsymbol{R}_k$. Hľadáme rozklad tohto priestoru na m disjunktných borelovských množín $A_1, \ldots, A_M$ tak, aby rozhodovanie bolo optimálne. A to v zmysle zaradenia vektoru do m-tej skupiny, pokiaľ $\boldsymbol{X}$ padne do množiny A_m.

Označme:

- s_{mj} stratu, ktorú utrpíme ak pozorovanie patriace do m-tej skupiny zaradíme do j-tej
- $f_m(\boldsymbol{X})$ hustotu rozdelenia vektoru $\boldsymbol{X}$ vzhľadom k σ konečnej miere μ, pre pozorovania patriace do skupiny K_m
- π_m pravdepodobnosť, že objekt patrí do skupiny K_m
- L ako strednú hodnotu straty.

Môžme vyjadriť

$$L = \sum_{m=1}^{M} \pi_m L_m, \tag{4.11}$$

kde L_m je podmienená stredná hodnota straty, pri podmienke, že pozorovanie patrí do skupiny m,

$$L_m = \sum_{j=1}^{M} s_{mj} \int_{A_j} f_m(\boldsymbol{X}) d\mu(\boldsymbol{X}).$$

Ak označíme $q_j(\boldsymbol{X}) = \sum_{m=1}^{M} \pi_m s_{mj} f_m(\boldsymbol{X})$, rovnica 4.11 bude tvaru

$$L = \sum_{j=1}^{M} \int_{A_j} q_j(\boldsymbol{X}) d\mu(\boldsymbol{X}).$$

Optimálnym rozkladom je pre nás ten, pre ktorý je stredná hodnota straty minimálna. Pokiaľ uvažujeme rozklad priestoru $\boldsymbol{R}_k$ na $A_1^{'}, \ldots, A_M^{'}$ tak, že platí

$$\boldsymbol{X} \in A_t^{'} \Rightarrow q_t(\boldsymbol{X}) \leq q_j(\boldsymbol{X}), j = 1, \ldots, M,$$

potom platí

$$\begin{aligned} L &= \sum_{j=1}^{M} \int_{A_j} q_j(\boldsymbol{X}) d\mu(\boldsymbol{X}) = \sum_{j=1}^{M} \sum_{t=1}^{M} \int_{A_j \cap A_t^{'}} q_j(\boldsymbol{X}) d\mu(\boldsymbol{X}) = \\ &= \sum_{t=1}^{M} \sum_{j=1}^{M} \int_{A_t^{'} \cap A_j} q_j(\boldsymbol{X}) d\mu(\boldsymbol{X}) \geq \sum_{t=1}^{M} \sum_{j=1}^{M} \int_{A_t^{'} \cap A_j} q_t(\boldsymbol{X}) d\mu(\boldsymbol{X}) = \\ &= \sum_{t=1}^{M} \int_{A_t^{'}} q_t(\boldsymbol{X}) d\mu(\boldsymbol{X}) = L^{'}. \end{aligned}$$

Často sa volí $s_{mm} = 0$ a $s_{mj} = 1$ pre $m \neq j$ a vlastne sa minimalizuje podiel chybne klasifikovaných pozorovaní. A teda

$$q_j(\boldsymbol{X}) = \sum_{m=1}^{M} \pi_m f_m(\boldsymbol{X}) - \pi_j f_j(\boldsymbol{X}).$$

Pri danom $\boldsymbol{X}$ potom platí

$$q_t(\boldsymbol{X}) \leq q_j(\boldsymbol{X}) \Leftrightarrow \pi_t f_t(\boldsymbol{X}) \geq \pi_j f_j(\boldsymbol{X}), j = 1, \ldots, M.$$

V prípade platnosti $\pi_t f_t(\boldsymbol{X}) > \pi_j f_j(\boldsymbol{X})$ pre všetky $j \neq t$, zaradíme pozorovanie do skupiny t. V Prípade rovnosti i pre iné j ako $j = t$ objekt zaradíme do ktorejkoľvek skupiny z maximálnych indexov.

V praxi sa tiež využíva predpoklad hustoty k-rozmerného normálneho rozdelenia so strednou hodnotu μ_j a variančnou maticou $\boldsymbol{\Sigma_j}$. Teda

$$f_j(\boldsymbol{X}) = (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma_j}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{X} - \mu_j)^T \boldsymbol{\Sigma_j}^{-1} (\boldsymbol{X} - \mu_j)\right\}.$$

Pre zjednodušenie výpočtu sa pracuje so zlogaritmovanou verziou nerovnosti

$$\ln \pi_t + \ln f_t(\boldsymbol{X}) > \ln \pi_j + \ln f_j(\boldsymbol{X}), j \neq t. \tag{4.12}$$

Dosadením hustoty do nerovnosti 4.12 dostaneme

$$\ln \pi_t - \frac{k}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma_t}| - \frac{1}{2} (\boldsymbol{X} - \mu_t)^T \boldsymbol{\Sigma_t}^{-1} (\boldsymbol{X} - \mu_t) > \\ > \ln \pi_j - \frac{k}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma_j}| - \frac{1}{2} (\boldsymbol{X} - \mu_j)^T \boldsymbol{\Sigma_j}^{-1} (\boldsymbol{X} - \mu_j).$$

Po substitúcii

$$D_j(\boldsymbol{X}) = \ln \pi_j - \frac{1}{2} \ln |\boldsymbol{\Sigma_j}| - \frac{1}{2} (\boldsymbol{X} - \mu_j)^T \boldsymbol{\Sigma_j}^{-1} (\boldsymbol{X} - \mu_j) \tag{4.13}$$

platí

$$D_t(\boldsymbol{X}) > D_j(\boldsymbol{X}), \qquad j \neq t. \tag{4.14}$$

$D_j(\boldsymbol{X})$ sa nazýva *kvadratická diskriminačná funkcia*. Pri praktickom výpočte sa určia všetky hodnoty $D_j(\boldsymbol{X})$, $j = 1, \ldots, M$ a pozorovanie zaradíme do skupiny t, ak platí $D_t(\boldsymbol{X}) = \max \{D_j(\boldsymbol{X}), j = 1, \ldots, M\}$.

V prípade, že $\boldsymbol{\Sigma_j} = \boldsymbol{\Sigma}$ pre všetky $j = 1, \ldots, M$, sa používa tzv. *lineárna diskriminačná funkcia*, pre ktorú platí

$$d_j(\boldsymbol{X}) = \ln \pi_j - \frac{1}{2} \mu_j^T \boldsymbol{\Sigma_j}^{-1} \mu_j + \mu_j^T \boldsymbol{\Sigma_j}^{-1} \boldsymbol{X}.$$

A nerovnosť 4.14 je ekvivalentná nerovnosti $d_t(\boldsymbol{X}) > d_j(\boldsymbol{X}), j \neq t$.

Pri výpočtoch sa tiež využívajú odhady $\overline{\boldsymbol{X}}_j$ ako odhad μ_j, $\boldsymbol{S}_j$ ako odhad variančnej matice $\boldsymbol{\Sigma_j}$. Odporúčaným odhadom pravdepodobnosti, že pozorovanie patrí do j-tej skupiny, je $\frac{1}{M}$ alebo relatívna početnosť, t. j. $\frac{n_j}{n}$.

Taktiež je vhodné overiť účinnosť diskriminačnej analýzy v prípadoch, keď už o určitých pozorovaniach vieme, do ktorej skupiny patria. Pre tieto pozorovania sa spočítajú hodnoty d^2_{sj} resp. D_j a je teda možné určiť, do ktorej skupiny by sme dané pozorovanie zaradili na základe kanonickej diskriminačnej analýzy resp. kvadratickej diskriminačnej funkcie. Týmto spôsobom je potom možné vyčísliť podiel správne a nesprávne klasifikovaných pozorovaní.

4.4 Aplikácia diskriminačnej analýzy

Popísanú metódu použijeme na analýzu našej dátovej matice zmlúv poistenia domácnosti. Aj v tejto časti budeme pracovať so softvérom *Wolfram Mathematica 8.0*, v ktorom úlohu spracujeme vytvorením vlastného kódu. Využijeme pritom výsledky z doterajších analýz a dátovú maticu budeme spracovávať bez premenných, ktoré sme už vylúčili pomocou analýzy hlavných komponentov, t. j. bez premenných `Cudzinec`, `Bydlisko` a `Titul`. Keďže naším cieľom je prepojiť skúmanie ziskovosti a podvodov poistných zmlúv, spojíme premenné `Profit` a `Fraud` do novej premennej `ProfitF`. Táto transformovaná premenná bude vyjadrovať či je daná zmluva zisková a zároveň bude zohľadňovať, či bol spáchaný poistný podvod.

K dispozícii teda máme 70 žiadostí rozdelených do dvoch skupín podľa premennej `ProfitF`, viď tabuľka 4.1.

Premenná	Kategórie	Popis
ProfitF	1	Zmluva je profitabilná. Za ziskové považujeme zmluvy, ktorých pôvodná premenná `Profit` bola v kategórii 1 alebo 2 a na zmluve nebol odhalený fraud.
	0	Zmluva nie je profitabilná. Teda pôvodná premenná `Profit` bola v kategórii 3 a 4 alebo bol na zmluve odhalený fraud.

Tabuľka 4.1: Transformovaná premenná `ProfitF`

Medzi neziskové zmluvy sme zaradili aj tie, ktoré sa na základe pôvodnej premennej `Profit` javili ako ziskové, ale zároveň bol na nich odhalený poistný podvod. Tieto zmluvy sú totiž pre poistiteľa z dlhodobého hľadiska stratové. Základom našej úvahy je fakt, že poisťovňa v prípade podvodu (ľubovoľnej výšky) podstupuje v každom prípade neoprávnenú výplatu poistného plnenia a tým inklinuje k nižšej profitabilite poistného kmeňa.

4.4.1 Test normality

Skôr než pristúpime k samotnej diskriminačnej analýze prevedieme test na overenie mnohorozmernej normality vstupných dát. Budeme pri tom vychádzať z mnohorozmernej šikmosti a špicatosti. Podľa [5] definujeme mnohorozmerný koeficient šikmosti ako

$$b_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T \boldsymbol{S}^{-1}(\boldsymbol{X}_j - \overline{\boldsymbol{X}})]^3,$$

kde $\boldsymbol{X}_i$ je k rozmerný vektor meraných znakov, n je počet pozorovaní, $\boldsymbol{S}$ je výberová kovariančná matica a $\overline{\boldsymbol{X}}$ je výberový priemer. Testujeme nulovú hypotézu, že dátová matica pochádza z mnohorozmerného normálneho rozdelenia. V rámci testu normality sa overuje nulová hypotéza o tom, že odchýlka od strednej hodnoty koeficientu šikmosti je nevýznamná, t. j. $E(b_1) = 0$. Túto hypotézu, ktorá značí mnohorozmerné normálne rozdelenie, zamietame na hladine α ak je hodnota testovej štatistiky

$$V = n\frac{b_1}{6}$$

väčšia ako kvantil $\chi^2(1-\alpha/2)$ chí-kvadrát rozdelenia s $k(k+1)(k+2)/6$ stupňami voľnosti.

Pre našu dátovú maticu, ktorú využijeme ako vstup pre diskriminačnú analýzu, platí $n = 70$ a $k = 13$. Koeficient šikmosti b_1 vychádza 40,6747 a hodnota štatistiky V je 474,538. Hodnota kvantilu $\chi^2(1-\alpha/2)$ so 455 stupňami voľnosti na hladine α =0,05 je 515,996. Nulovú hypotézu teda zamietnuť nemôžme.

Ďalej definujeme mnohorozmerný koeficient špicatosti ako

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T \boldsymbol{S}^{-1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})]^2.$$

Testuje sa nulová hypotéza, že odchýlka od strednej hodnoty koeficientu špicatosti je významná, t. j. $E(b_2) = k(k+2)$. A teda hypotézu o normalite mnohorozmerného rozdelenia zamietame na hladine významnosti α, pokiaľ prekročí hodnota testovej štatistiky

$$U = n\sqrt{\frac{n}{8k(k+2)}}(b_2 - k(k+2))$$

kvantil normovaného normálneho rozdelenia $u_{1-\alpha/2}$.

Mnohorozmerná špicatosť pre našu dátovú maticu vychádza 187,516. Testová štatistika U má hodnotu -1,58526, čo je nižšia hodnota ako kvantil $u_{0,975}$ =1,95996 normovaného normálneho rozdelenia pre α =0,05. Opäť teda nie je možné zamietnuť hypotézu a mnohorozmernej normalite dátovej matice. Odhady mnohorozmerných koeficientov šikmosti a špicatosti sme taktiež spočítali v *Mathematice 8.0* a kód prikladáme v Prílohe C.1.

4.4.2 Diskriminačná analýza v praxi

Prejdeme teraz k samotnej diskriminačnej analýze a začneme kanonickou diskriminačnou analýzou. Ako sme uviedli v úvode tejto podkapitoly, dáta máme rozdelené do 2 skupín. Našim prvým zámerom je nájsť premenné, ktoré majú vplyv na odlíšenie zmlúv, ktoré sú zaradené v skupine profitabilných od skupiny neprofitabilných zmlúv.

Algoritmus rozdelíme na niekoľko častí, v ktorých uvedieme postup výpočtu a výsledné hodnoty. Kód z *Mathematici* uvádzame v Prílohe C.2.

- *Vstupné parametre:* vstupom do *Mathematici* bude upravená dátová matica bez premenných `Cudzinec`, `Bydlisko` a `Titul`, počet skupín 2, počet meraných znakov 13 a určíme tiež počet pozorovaní v jednotlivých skupinách, odlíšených podľa premennej `ProfitF`. V skupine profitabilných zmlúv (značených 1) máme zaradených 46 pozorovaní a zvyšných 24 považujeme za neprofitabilné.

- *Vektory priemerných hodnôt:* Spočítame priemer jednotlivých meraných znakov cez všetky pozorovania a taktiež priemery znakov cez pozorovania v jednotlivých skupinách. Priemery cez všetky pozorovania, zaokrúhlené na 3 desatinné miesta uvádzame pre prehľadnosť aj s názvom premennej:

$$\overline{X} = \begin{pmatrix} \text{Stvrt} & \text{Mesto} & \text{TO} & \text{PC} & \text{Vek} & \text{Pohlavie} & \text{Adr} & \text{SP} & \text{Platenie} & \text{BL} & \text{DalsieP} & \text{Zlavy1} & \text{Zlavy2} \\ 2,071 & 3,257 & 1,171 & 3,1 & 41,271 & 1,329 & 1,386 & 1,8 & 1,471 & 1,829 & 1,9 & 1,514 & 1,371 \end{pmatrix}^T$$

  Priemery v skupinách s označením $\overline{X}_{(1)}$ pre skupinu profitabilných zmlúv a $\overline{X}_{(0)}$ pre skupinu neprofitabilných zmlúv sú potom nasledovné:

$$\overline{X}_{(1)} = (1,783; 2,978; 1,109; 3,609; 42,413; 1,348; 1,261; 1,804; 1,304; 1,891; 1,826; 1,63; 1,478)^T$$

$$\overline{X}_{(0)} = (2,625; 3,792; 1,292; 2,125; 39,083; 1,292; 1,625; 1,792; 1,792; 1,708; 2,042; 1,292; 1,167)^T$$

- *Výpočet potrebných matíc:* V tejto časti spočítame výberové kovariančné matice pre obe naše skupiny pozorovaní. Určené matice uvádzame v Prílohe C.3 s hodnotami zaokrúhlenými na 4 desatinné miesta.

  K ďalším výpočtom budeme potrebovať aj maticu medziskupinovej variability $\boldsymbol{B}$, ktorú spočítame podľa 4.2, maticu vnútroskupinovej variablity $\boldsymbol{E}$, ktorú spočítame podľa 4.1, a k nej diagonálnu maticu $\boldsymbol{H}$. Všetky spomínané matice sú uvedené v Prílohe C.4.

- *Riešenie sústavy 4.4:* V tomto kroku vyriešime sústavu rovníc s cieľom získať vlastné čísla matice $(\boldsymbol{B} - \lambda\boldsymbol{E})$. Využijeme k tomu normovaciu podmienku 4.6, aby sme získali koeficienty $\boldsymbol{c}$, ktoré použijeme k ďalším výpočtom. Keďže máme len dve skupiny, ktoré sa snažíme odlíšiť, riešením sústavy je jedno vlastné číslo rovné 0,921738. Za zvolenej normovacej podmienky tomuto vlastnému číslu náležia 2 vlastné vektory:

$$\begin{aligned}\boldsymbol{c}_1 &= (0,41796; 0,109732; 0,618474; -0,437984; -0,00170564; 0,136913; 0,374419;\\ &\quad -0,107271; 0,690459; -0,907722; -0,0558716; -0,30854; -0,187622)^T,\\ \boldsymbol{c}_1^{'} &= (-0,41796; -0,109732; -0,618474; 0,437984; 0,00170564; -0,136913;\\ &\quad -0,374419; 0,107271; -0,690459; 0,907722; 0,0558716; 0,30854; 0,187622)^T.\end{aligned}$$

  Ďalej nám postačia absolútne hodnoty zložiek jednotlivých vektorov. Použijeme teda už len prvý z uvedených vektorov ako vektor $\boldsymbol{c}_1$.

- *Výpočet koeficientov:* Z rovnice 4.7 spočítame priemerné hodnoty diskriminantov v oboch skupinách, t. j. $\overline{y}_{11}$ = -0,683496 a $\overline{y}_{21}$ = 1,31003, kde nám konštatna v_1 ako súčin celkového priemeru $\overline{\boldsymbol{X}}$ a vektoru $-\boldsymbol{c}_1$ vyšla 0,447366. Teraz už máme všetko potrebné na vyjadrenie normovaných koeficientov vektoru $\boldsymbol{c}_1$ z rovnice 4.8 a korelačných koeficientov z 4.9 medzi kanonickou premennou a pôvodnými premennými. Hodnoty sú uvedené v tabuľke 4.2.

Premenná	**Normované koeficienty**	**Korelačné koeficienty**
Stvrt	0,275062	0,642089
Mesto	0,190498	0,235033
TO	0,23013	0,246665
PC	-0,430124	-0,757855
Vek	-0,0201846	-0,141141
Pohlavie	0,0651409	-0,0592095
Adr	0,172865	0,395627
SP	-0,04353	-0,0156758
Platenie	0,456293	0,369901
BL	-0,337758	-0,246665
DalsieP	-0,0548629	0,110128
Zlavy1	-0,14814	-0,353933
Zlavy2	-0,0875649	-0,334904

Tabuľka 4.2: Normované a korelačné koeficienty

Na základe normovaných koeficientov sa ako menej významné znaky na odlíšenie ziskových a neziskových zmlúv javia vek, pohlavie, súčasné poistenie u iného poistiteľa, ďalšie poistenie a zľavy na prírodné živly t. j. premenné `Vek, Pohlavie, SP, DalsieP, Zlavy2`. Ako kritérium sme brali absolútnu hodnotu normovaných koeficientov vo výške 0,1. Z vektoru korelačných koeficientov, kde sme ako kritérium volili absolútnu hodnotu koeficientu vo výške 0,15, môžme usúdiť zhodu v premenných, ktoré vychádzajú ako nepodstatné na rozlíšenie našich skupín až na premennú `Zlavy2`. Táto premenná sa javí ako významná. Aj kvôli hraničnej hodnote v prípade kritéria normovaných koeficientov sme sa rozhodli v našej ďalšej analýze túto premennú ponechať. Myslíme si totiž, že stanovovanie zliav v praxi často robí konkrétny produkt pre klienta atraktívnejším a výška celkovej zľavy môže z pohľadu poistiteľa významne ovplyvniť ziskovosť danej zmluvy.

V tento moment vieme, ktoré z premenných sú dôležité pre klasifikáciu profitabilných skupín. V rámci tvorby prediktívneho modelu sa tak dostávame k ďalšej časti, a to zaradeniu nových pozorovaní - nových žiadostí o poistenie domácnosti. Využijeme pritom maticu vzdialeností kanonickej diskriminačnej analýzy podľa 4.10 a taktiež maticu kvadratických diskriminačných funkcií na základe rovnice 4.13.

Najprv aplikáciou oboch metód zaradíme pozorovania obsiahnuté v analyzovanom súbore. Pomocou tejto klasifikácie budeme môcť stanoviť úspešnosť určených modelov, t. j. budeme schopní odhadnúť ako úspešne budú klasifikovať nové žiadosti.

K ďalším výpočtom však budeme potrebovať niektoré predchádzajúce výpočty, preto najprv aplikujeme kanonickú analýzu na našu dátovú maticu s menším počtom meraných znakov. V Prílohe C.5 uvádzame potrebné výsledky upravenej dátovej matice. Nenulové vlastné číslo upravenej matice je opäť len jedno 0,914982. Jeho príslušný vlastný vektor je

$$\boldsymbol{c}_1^1 = (-0,425258; -0,105316; -0,638908; 0,43131, -0,330086; -0,676185; 0,874973; \\ 0,324416; 0,21107)^T,$$

priemerné hodnoty diskriminantov v oboch skupinách sú $\overline{y}_{11}^1$ =0.680986, $\overline{y}_{21}^1$=-1.30522 a konštanta v_1^1 =-0.293017. Ani v tomto prípade sme na základe mnohorozmerného koeficientu šikmosti a špicatosti nezamietli hypotézu o normalite mnohorozmerného rozdelenia dátovej matice.

Z určenej matice vzdialeností pozorovaní od jednotlivých zhlukov, ktorú uvádzame v Prílohe C.6, vyplýva, že náš model kanonickej diskriminačnej analýzy chybne zaradí zmluvy s poradovým číslom 9, 10, 17, 24, 31, 32, 35, 36, 38, 47, 49, 55, 57. Celkovo 13 zmlúv a úspešnosť kanonickej diskriminačnej analýzy je teda $100\frac{57}{70} \approx 81{,}43\%$.

Pre klasifikáciu pomocou kvadratických diskriminačných funkcií použijeme ako odhad pravdepodobnosti príslušnosti k jednotlivým skupinám relatívnu početnosť v celom súbore pozorovaní. V tomto prípade budú podľa hodnôt v matici kvadratických funkcií uvedených v Prílohe C.6 nesprávne klasifikované zmluvy s poradovými číslami 9, 10, 35, 36, 47, 49, 55, 57. Celkovo 8 zmlúv a úspešnosť metódy kvadratických diskriminačných funkcií je teda $100\frac{62}{70} \approx 88{,}57\%$.

V oboch prípadoch tak môžme odôvodnene predpokladať, že naše modely budú dostatočne efektívne k zaradeniu nových pozorovaní, teda predikcii profitability nových žiadostí o poistenie. Pre ilustráciu zaradenia nových pozorovaní využijeme 12 náhodne vybraných nových žiadostí, ktoré sú uvedené v tabuľke 4.3. Pre tieto žiadosti uvádzame hodnoty 9 premenných, ktoré boli označené ako podstatné pre klasifikáciu profitabilných zmlúv.

Pozorovanie	Stvrt	Mesto	TO	PC	ADR	Platenie	BL	Zlavy1	Zlavy2
71.	2	2	1	4	1	2	2	1	2
72.	1	1	1	5	1	3	2	2	1
73.	2	4	2	1	1	2	1	1	1
74.	2	3	1	4	1	1	2	1	1
75.	3	5	1	4	1	1	2	2	1
76.	1	3	1	5	2	1	2	1	2
77.	1	6	2	3	1	2	2	2	2
78.	3	5	1	2	2	2	2	2	2
79.	3	3	1	4	2	1	2	2	2
80.	2	2	1	2	2	2	1	2	1
81.	2	5	2	4	2	1	2	1	1
82.	3	1	1	2	1	3	2	1	2

Tabuľka 4.3: Nové žiadosti

Matica vzdialeností diskriminačnej analýzy a kvadratických diskriminačných funkcií je uvedená v Prílohe C.7. Podľa týchto výsledkov uvádzame klasifikáciu nových žiadostí v tabuľke 4.4. Hodnoty klasifikácie vyjadrujú predikciu premennej ProfitF, t. j. rozdelenie na profitabilné zmluvy (hodnota 1) a neprofitabilné zmluvy (hodnota 0).

Pozorovanie	**Klasifikácia na základe kanonickej DA**	**Klasifikácia na základe kvadratických diskriminačných funkcií**
71.	1	1
72.	1	1
73.	0	0
74.	1	1
75.	1	1
76.	1	1
77.	1	1
78.	0	0
79.	1	0
80.	0	0
81.	1	1
82.	0	0

Tabuľka 4.4: Klasifikácia nových žiadostí

Z tabuľky klasifikácie nových pozorovaní vidíme, že medzi neziskové by sme na základe oboch metód zaradili pozorovania 73, 78, 80 a 82. Z nich sa dve pozorovania vyskytli na internom black liste (73 a 80) a zvyšné dve sa nachádzajú v rizikovej oblasti. Za povšimnutie stojí fakt, že ani jedno z týchto pozorovaní nemá poistnú čiastku vyššiu ako 750 tisíc Kč. Výsledok klasifikácie je rôzny v prípade pozorovania 79. Z parametrov tejto žiadosti vidíme, že sa jedná o rizikovú oblasť s vyššou poistnou čiastkou. Môžme teda usudzovať, že v prípade poistnej udalosti hrozí, že by bolo vyplatené poistné plnenie, ktoré by zmluvu zaradilo skôr do skupiny neprofitabilných zmlúv.

Kapitola 5

Logistická regresia

5.1 Základné charakteristiky

V tejto kapitole sa pokúsime, pre porovnanie s výsledkami diskriminačnej analýzy, nájsť regresný model, ktorý by na základe poskytnutých informácií o klientovi umožnil odhadnúť pravdepodobnosť ziskovosti danej poistnej zmluvy. Využijeme pri tom logistickú regresiu. Medzi mnohými štatistikmi prevláda názor, že logistická regresia je univerzálnejšia a vhodnejšia pre väčšinu situácií než diskriminačná analýza. Jej výhodou oproti diskriminačnej analýze je fakt, že nekladie žiadne podmienky na tvar rozdelenia a nevyžaduje ani predpoklady na apriórne pravdepodobnosti. Vybudovanie modelu logistickej regresie popíšeme teoreticky, a podobne ako v predchádzajúcej kapitole, ho následne využijeme na odhad profitability nových žiadostí poistenia domácností.

Stále uvažujme pre i-teho klienta vektor k dimenzionálneho priestoru $\boldsymbol{R}_k$ meraných znakov $\boldsymbol{x}_i, i = 1, \ldots, n$. Binárna vysvetľovaná premenná Y_i nadobúda hodnôt 1 a 0 podľa zaraďenia pozorovaní do kategórie. Ďalej označme $\pi(\boldsymbol{x}_i) = P(Y_i = 1|\boldsymbol{x}_i)$. Všimnime si tiež, že ide zároveň o strednú hodnotu Y_i,

$$E(Y_i|\boldsymbol{x}_i) = 1 * P(Y_i = 1|\boldsymbol{x}_i) + 0 * P(Y_i = 0|\boldsymbol{x}_i) = P(Y_i = 1|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i).$$

Túto pravdepodobnosť určíme pomocou vektoru charakteristík $\boldsymbol{x}_i$ i-teho klienta. Keďže ide o pravdepodobnosť, ktorej hodnoty ležia v intervale $\langle 0, 1\rangle$, použijeme logitovú transformáciu s vektorom neznámych parametrov $\boldsymbol{\beta}$. Podľa [6] predpokladáme, že $\pi(\boldsymbol{x}_i)$ má tvar

$$\pi(\boldsymbol{x}_i) = \frac{exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}. \tag{5.1}$$

Pričom funkcia $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$ sa nazýva *logit* a pomer pravdepodobností $\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}$ sa nazýva *šanca* respektíve *odds*. Jednoduchou úpravou je potom možné dospieť k tomu, že $\log(odds) = logit$.

5.2 Odhad parametrov modelu

Predpokladáme, že máme n nezávislých pozorovaní vektoru $(y_i, \boldsymbol{x}_i)$, pre jednotlivé $i = 1 \ldots, n$. Odhad parametrov vektoru $\boldsymbol{\beta}$ prevedieme pomocou metódy

maximálnej vierohodnosti, viď [2]. Táto metóda je založená na konštrukcii *vierohodnostnej funkcie* a hľadaní jej maxima vzhľadom k množine neznámych parametrov. Všeobecne možno pre i-te pozorovanie vyjadriť pravdepodobnosť

$$P(Y_i = y_i) = \pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{1-y_i}.$$

Platí $Y_i \sim Alt(\pi(\boldsymbol{x}_i))$, t. j.

$$P(Y_i = 1) = \pi(\boldsymbol{x}_i), \qquad P(Y_i = 0) = 1 - \pi(\boldsymbol{x}_i).$$

Keďže pozorované hodnoty sú podľa predpokladu nezávislé, dostaneme vierohodnostnú funkciu $L(\boldsymbol{\beta})$ ako súčin pravdepodobností pre jednotlivé pozorovania tvaru

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{1-y_i}. \tag{5.2}$$

K nájdeniu maxima funkcie 5.2 využijeme možnosť zlogaritmovania vierohodnostnej funkcie $L(\boldsymbol{\beta})$. Tento postup je často využívaný najmä z dôvodu zjednodušenia funkcie pre potreby následného derivovania a zároveň touto transformáciou nie je ovplyvnená poloha hľadaného extrému. Dostávame teda logaritmickú vierohodnostnú funkciu v tvare

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(\pi(\boldsymbol{x}_i)) + (1 - y_i) \log(1 - \pi(\boldsymbol{x}_i))].$$

Do tohto vzťahu dosadíme $\pi(\boldsymbol{x}_i)$ podľa 5.1. Funkciu derivujeme podľa vektoru parametrov $\boldsymbol{\beta}$ a pomocou retiazkového pravidla získame parciálne derivácie

$$\frac{\partial \log L(\beta_0)}{\partial \beta_0} = \frac{\partial \log L(\beta_0)}{\partial \pi(\boldsymbol{x}_i)} . \frac{\partial \pi(\boldsymbol{x}_i)}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \pi(\boldsymbol{x}_i)),$$

$$\frac{\partial \log L(\beta_j)}{\partial \beta_j} = \frac{\partial \log L(\beta_j)}{\partial \pi(\boldsymbol{x}_i)} . \frac{\partial \pi(\boldsymbol{x}_i)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - \pi(\boldsymbol{x}_i)),$$

pre $j = 1, 2, \ldots, k$, kde x_{ij} je j-tá zložka vektoru $\boldsymbol{x}_i$. Rovnice položíme rovné nule a dostaneme sústavu $k + 1$ vierohodnostných rovníc. Riešením sústavy získame maximálne vierohodnný odhad $\hat{\boldsymbol{\beta}}$ vektoru $\boldsymbol{\beta}$. Sústava sa rieši numericky napríklad pomocou *Newton-Raphsonovho iteračného algoritmu*. Tento algoritmus začína ľubovoľne zvolenou hodnotou $\boldsymbol{\beta}^{(0)}$ a ďalej postupuje iteračne

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{q}^{(t)}, \tag{5.3}$$

kde $\boldsymbol{\beta}^{(t)}$ je aktuálny odhad vektoru parametrov. $\boldsymbol{H}^{(t)}$ je matica druhých parciálnych derivácií logaritimickej vierohodnostnej funkcie $logL$ v bode $\boldsymbol{\beta}^{(t)}$, ktorej prvky sú tvaru

$$h_{ab}^{(t)} = \sum_{i=1}^{n} x_{ia}x_{ib}(\pi^{(t)}(\boldsymbol{x}_i)(1 - \pi^{(t)}(\boldsymbol{x}_i))), \qquad a, b = 1, 2, \ldots, k$$

a $\boldsymbol{q}^t$ je vektor prvých parciálnych derivácií $logL$ v bode $\boldsymbol{\beta}^{(t)}$ tvaru

$$q_j^{(t)} = \sum_{i=1}^{n} x_{ij}(y_i - \pi^{(t)}(\boldsymbol{x}_i), \qquad j = 1, 2, \ldots, k,$$

kde

$$\pi^{(t)}(\boldsymbol{x}_i) = \frac{exp(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)})}{1 + exp(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)})}, \qquad t = 0, 1, \ldots.$$

Postupne dostaneme lepšie odhady parametrov $\boldsymbol{\beta}$ a platí $\boldsymbol{\beta}^{(t)} \to \widehat{\boldsymbol{\beta}}$, pre $t \to \infty$. Maximálne vierohodným odhadom $\pi(\boldsymbol{x}_i)$ pre $i = 1, \ldots, n$ je teda

$$\widehat{\pi}(\boldsymbol{x}_i) = \frac{exp(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})}{1 + exp(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})}.$$

Z vlastností maximálne vierohodných odhadov, viď [2], plynie, že odhad $\widehat{\boldsymbol{\beta}}$ má asymptoticky normálne rozdelenie so strednou hodnotou $\boldsymbol{\beta}$ a rozptylom, ktorý je rovný inverzii Fisherovej informačnej matice $\boldsymbol{J}(\boldsymbol{\beta})$, ktorá je v našom prípade pre veľké t rovná matici $\boldsymbol{H}^{(t)}$ a má tvar

$$\boldsymbol{J}_{ab} = -\sum_{i=1}^{n} x_{ia} x_{ib} \pi(\boldsymbol{x}_i)(1 - \pi(\boldsymbol{x}_i)), \qquad a, b = 1, 2, \ldots, k.$$

Platí teda $var(\boldsymbol{\beta}) = \boldsymbol{J}^{-1}(\boldsymbol{\beta})$. Rozptyl a-tej zložky vektoru $\boldsymbol{\beta}$ je a-ty diagonálny prvok tejto matice, t. j. $var(\beta_a)$. Kovarianciu zložiek $cov(\beta_a \beta_b)$ získame ako mimo diagonálny prvok tejto matice. Odhad rozptylu $\widehat{var}(\widehat{\boldsymbol{\beta}})$ a kovariancie dostaneme, ak v spočítame maticu $var(\boldsymbol{\beta})$ s maximálne vierohodným odhadom parametrov $\widehat{\boldsymbol{\beta}}$. Podľa [6] pre informačnú maticu platí

$$\boldsymbol{J}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X},$$

kde $\boldsymbol{X}$ je matica $k + 1$ regresorov

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix},$$

a $\boldsymbol{V}$ je diagonálna matica tvaru

$$\boldsymbol{V} = \begin{pmatrix} \pi(\boldsymbol{x}_1)(1 - \pi(\boldsymbol{x}_1)) & 0 & \ldots & 0 \\ 0 & \pi(\boldsymbol{x}_2)(1 - \pi(\boldsymbol{x}_2)) & \ldots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \ldots & 0 & \pi(\boldsymbol{x}_n)(1 - \pi(\boldsymbol{x}_n)) \end{pmatrix}.$$

5.3 Selekcia premenných a konštrukcia modelu

Pre konštrukciu modelu logistickej regresie je dôležitý aj výber a ohodnotenie významnosti jednotlivých parametrov. Testom pomerom vierohodností (Wilksov test) sa overuje hypotéza o nulovosti niektorého podvektoru vektora $\boldsymbol{\beta}$. Tento test je založený na vierohodnostnom pomere

$$LR = -2\log\left(\frac{\text{vierohodnosť zjednodušeného modelu}}{\text{vierohodnosť modelu s nenulovými parametrami}}\right) \tag{5.4}$$

Za platnosti hypotézy, že zložky $i_1, i_2, \ldots, i_l$ vektoru $\boldsymbol{\beta}$ sú nulové, má štatistika LR asymptoticky rozdelenie χ^2 o $k-l$ stupňoch voľnosti. Hypotézu zamietame, ak je hodnota štatistiky 5.4 väčšia ako príslušný horný kvantil rozdelenia χ^2.

Po otestovaní modelu na nenulovosť aspoň jedného parametru $\boldsymbol{\beta}$ je vhodné previesť orientačný test na významnosť jednotlivých parametrov tzv. *Waldov test*. Pre *Waldov test* sa overuje nulová hypotéza, že parameter $\beta_j = 0$, proti alternatíve $\beta_j \neq 0$. Definuje sa pritom veličina

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{var(\hat{\beta}_j)}} \overset{as}{\sim} N(0,1),$$

kde $var(\hat{\beta}_j)$ získame z odhadu inverznej Fisherovej informačnej matice. Za platnosti nulovej hypotézy má táto veličina asymptoticky normálne rozdelenie s nulovou strednou hodnotou a jednotkovým rozptylom. *Waldovu štatistiku* potom definujeme ako $W_j = Z_j^2$, ktorá má za platnosti nulovej hypotézy asymptoticky χ^2 rozdelenie s jedným stupňom voľnosti. Nulovú hypotézu zamietame, ak $W_j > \chi_1^2(\alpha)$, kde $\chi_1^2(\alpha)$ je α kvantil χ_1^2 rozdelenia. Pomocou tohto testu vieme určiť, ktoré parametre sú pre model významné, a ktoré sú naopak kandidátmi pre nulovosť.

Pre testovanie hypotézy, že všetky parametre sú nulové sa využíva mnohorozmerná varianta Waldovej testovej štatistiky

$$\boldsymbol{W} = \hat{\boldsymbol{\beta}}^T[\widehat{var}(\hat{\boldsymbol{\beta}})]^{-1}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}\hat{\boldsymbol{\beta}},$$

ktorá má asymptoticky rozdelenie χ^2 o k stupňoch voľnosti. Detailne pozri napr. [2] alebo [6].

Jedným z prístupov pre voľbu signifikantných premenných a konštrukciu modelu je tzv. *stepwise metóda*. Pomocou tejto metódy sú premenné na základe štatistických kritérií vybrané buď pre zahrnutie, alebo vylúčenie z modelu sekvenčným spôsobom. Rozlišujú sa dve postupné konštrukcie modelu:

- *Spätná regresia* (backward elimination), pri ktorej sa vychádza z plného modelu a postupne sa podľa významnosti znižuje počet parametrov.
- *Rozširujúca sa regresia* (forward selection), kedy sa vychádza z modelu, ktorý obsahuje len jeden parameter a postupne sa premenné do modelu pridávajú.

Stručne popíšeme model spätnej regresie. Základný model by sme mohli zhrnúť do nasledujúcich krokov:

1. **Krok 0**: Nech $M \subseteq \{1, 2, \ldots, k\}$ je množina indexov parametrov vektoru $\boldsymbol{\beta}$. Označme L_M hodnotu logaritmu vierohodnostnej funkcie za predpokladu, že v modeli sú parametre s indexom z množiny M. Ďalej predpokladáme, že v modeli sú zahrnuté všetky parametere. Vylúčenie parametra je založené na významnosti štatistiky

$$G_j^{(0)} = 2(L_{\{1,2,\ldots,k\}} - L_{\{1,2,\ldots,k\}\setminus\{j\}}), \tag{5.5}$$

   ktorá má za platnosti $H_0 : \beta_j = 0$ rozdelenie χ^2 o jednom stupni voľnosti. Označme ďalej $p_j^{(0)} = P(\chi_1^2 > G_j^{(0)})$. Čím je táto pravdepodobnosť menšia, tým je model presnejší. Vyberieme teda index $r_0 = \max p_j^{(0)}$, najväčšej hodnoty $p^{(0)}$. Pokiaľ bude $p_{r_0} < p_R$ pre predom stanovenú hraničnú hodnotu (viď popis nižšie), algoritmus skončí. Inak položíme index iterácie $s = 0$, $R_0 = M \setminus \{r_0\}$ a pokračujeme krokom 1.

2. **Krok 1**: Označíme R_s množinu indexov po s-tej iterácii. Položme $s = s+1$. Vyberieme rovnakým spôsobom novú premennú, ktorú vylúčime z modelu. Určíme teda pre každé $j \in R_{s-1}$:

$$G_j^{(s)} = 2(L_{R_{s-1}} - L_{R_{(s-1)\setminus\{j\}}}).$$

   Ďalej určíme hodnotu $r_s = \max p_j^{(s)}$ pre $j \in R_{s-1}$, $p_j^{(s)} = P(\chi_1^2 > G_j^{(s)})$. Pokiaľ $p_{r_s} \geq p_R$ premennú s indexom r_s vylúčime z modelu, teda $R_s = R_{s-1} \setminus \{r_s\}$ a opakujeme krok 1. Algoritmus skončí v prípade $p_{r_s} < p_R$. Konečnosť algoritmu je zaručená konečnosťou počtu premenných v modeli.

Po tomto kroku ešte môže nasledovať ďalší krok, v ktorom sa overuje prípadné opätovné pridanie už vylúčenej premennej do modelu.

Významnú úlohu v algoritme spätnej i rozširujúcej sa regresie hrajú hraničné hodnoty pre zahrnutie alebo vylúčenie premennej p_E (pre rozširujúcu sa regresiu) a p_R[1] (pre spätnú regresiu podľa uvedeného algoritmu). Podľa [6] je vhodné voliť hodnoty $p_E \in (0,15; 0,20)$ a hodnoty $p_R = p_E + (0,05; 0,01)$. Ukazuje sa totiž, že hodnoty kritickej hladiny z teórie štatistických hypotéz sú príliš obmedzujúce. Pokiaľ chceme zaručiť, aby model obsahoval viac premenných, volíme p_E ešte väčšie a rovnako tak vyššou hodnotou p_R zaručíme malé vylučovanie zaradených premenných do modelu.

5.4 Diverzifikačná schopnosť modelu

Schopnosť diverzifikácie modelu udáva nakoľko je model schopný rozlíšiť pozorovania, ktorých odozva modelu je rovná jednej, od tých s odozvou rovnej nule. Uvažujme model, ktorým sa snažíme pre každé pozorovanie odhadnúť hodnotu

[1] Indexy E a R sú odvodené zo slov *Enter* ako vstúpiť rep. *Remove* ako vylúčiť.

y_i binárnej premennej Y_i, $i = 1, \ldots, n$. Každému pozorovaniu je priradené na základe meraných znakov skóre $s_i \in R$. Ako skóre sa používa napríklad odhad pravdepodobnosti $\widehat{\pi}(\boldsymbol{x}_i)$ z modelu logistickej regresie. Cieľom je potom stanoviť hranicu s_0 tak, aby všetky pozorovania, ktorých skóre je vyššie ako s_0, mali odozvu modelu rovnú jednej. Schopnosť diverzifikácie modelu teda hovorí o tom ako je model na základe stanoveného skóre schopný rozlíšiť pozorovania s $y_i = 1$ od tých s $y_i = 0$.

K ukazovateľom diverzifikácie patrí tzv. *Giniho koeficient*, ktorý vychádza z *Lorenzovej krivky*, označovanej tiež *ROC krivka*. Na zostrojenie tejto krivky sa využívajú odhady distribučných funkcií skóre pre pozorovania s $y_i = 1$ a $y_i = 0$. Empirické distribučné funkcie sú tvaru

$$F_1(s) = \frac{1}{n_1} \sum_{j=1}^{n} I_{(-\infty,s)}(s_j) y_j, s \in R,$$

$$F_0(s) = \frac{1}{n_0} \sum_{j=1}^{n} I_{(-\infty,s)}(s_j)(1 - y_j), s \in R,$$

kde n_1 vyjadruje počet pozorovaní s $y_i = 1$, teda $n_1 = \sum_{j=1}^{n} y_i$ a $n_0 = n - n_1$. Hodnota $I_{(-\infty,s)}(s_j)$ je rovná 1, pokiaľ $s_j \in (-\infty, s]$, inak je hodnota nulová. Lorenzova krivka potom vzniká spojením bodov $[F_0(s_i), F_1(s_i)]$, $i = 1, \ldots, n$. Krajnými bodmi sú $[0, 0]$, $[1, 1]$, a teda krivka leží vnútri jednotkového štvorca. Pokiaľ krivka leží bližšie k stranám štvorca, má model lepšiu diverzifikačnú schopnosť ako keď krivka leží bližšie k diagonále. Bod, ktorý leží najbližšie k ľavému hornému okraju, sa označuje ako hraničný bod pre klasifikáciu do jednotlivých skupín.

Možnosťou merania diverzifikačnej sily modelu je potom spočítanie plochy pod krivkou, označovanej ako AUC, viď [6]. Na jej základe je založený aj spomínaný Giniho koeficient. Pokiaľ označíme A plochu medzi Lorenzovou krivkou a diagonálou, B ako plochu nad Lorenzovou krivkou, platí, že $B + AUC = 1$ (plocha celého jednotkového štvorca). Potom $A + B = \frac{1}{2}$ a Giniho koeficient je definovaný ako pomer plochy A k ploche nad diagonálou, t. j.

$$Gini = \frac{A}{A + B} = 2A = 1 - 2B = 1 - 2(1 - AUC) = 2AUC - 1.$$

Platí $Gini \in \langle -1, 1 \rangle$, kde hodnoty bližšie k 1 naznačujú lepšiu diverzifikačnú schopnosť, v prípade hodnoty 0 model nemá žiadnu diverzifikačnú silu a záporné hodnoty hovoria o opačne postavenom modeli.

5.5 Aplikácia na dáta

V ďalšej časti budeme hľadať model logistickej regresie, ktorý by bolo možné využiť na klasifikáciu nových žiadostí do predom stanovených skupín, profitabilných resp. neprofitabilných zmlúv. Využijeme k tomu štatistický softvér *NCSS 2007*. Ako vstupné dáta pre nastavenie modelu budeme používať rovnakú dátovú maticu, ktorú sme použili v časti 4.4. Teda dátovú maticu bez premenných

Cudzinec, Bydlisko a Titul. Vysvetľovanou premennou bude premenná ProfitF a vysvetľujúcimi premennými sú merané znaky Stvrt, Mesto, TO, PC, Vek, Pohlavie, Adr, SP, Platenie, BL, DalsieP, Zlavy1 a Zlavy2.

5.5.1 Kódovanie premenných

Keďže máme len jednu spojitú numerickú premennú Vek a ostatné sú kategoriálne (nominálne i ordinálne), je vhodné premenné zakódovať umelými premennými. Hosmer a Lemshow, viď [6], nazvali tieto premenné *designové*. V logistickej regresii totiž nie je vhodné kombinovať rovnakým spôsobom číselné a kategoriálne premenné. Voľbou vhodných designových premenných by sme sa mali zaoberať v prípadoch, keď potrebujeme porovnať úroveň vysvetľovanej premennej pri zmene kategórie tej premennej, ktorá nemá číselný význam. Používanie týchto umelých premenných je teda úzko spojené s interpretáciou jednotlivých parametrov. Výber vhodného kódovania by sa malo odvíjať od typu premennej. Pre rôzne typy designových premenných získame rôzne hodnoty parametrov $\boldsymbol{\beta}$.

My sme zvolili kódovanie premennými, často označovanými ako *dummy* premenné. Toto kódovanie je najpoužívanejšie, zahrnuté vo väčšine štatistických softvéroch a jednoduchšie na interpretáciu. Ide o kódovanie premenných pomocou 0 a 1, pričom sa zvolí jedna referenčná kategória, ku ktorej sa vzťahujú všetky ostatné. Pre referenčnú kategóriu sú všetky dummy premenné nulové. Každá ďalšia kategória obsahuje jednu dummy premennú rovnú 1. Na zakódovanie premennej, ktorá má k kategórií, je potrebných $k-1$ dummy premenných. V našom prípade sme za referenčnú kategóriu volili vždy tú prvú. Kódovanie premenných uvádzame v Prílohe D.1. Na základe tohto kódovania pracujeme ďalej s novými premennými, ktorých pomenovanie je doplnené o index vyjadrujúci ich sekvenčné poradie.

5.5.2 Vybudovanie modelu

Pristúpime teraz k samotnému odhadu modelu logistickej regresie pre náš dátový súbor. K vybudovaniu modelu využijeme zabudovanú procedúru, ktorú ponúka program *NCSS* pre odhad modelu logistickej regresie. Budeme vytvárať model, v ktorom je každá premenná využívaná samostatne a nie sú použité interakcie ani žiadne dodatočné podmienky medzi premennými. V podstate budeme postupovať spätnou regresiou tak, ako bola definovaná v časti 5.3. Začneme s odhadom modelu, ktorý zahŕňa všetky premenné a postupnými iteráciami vylúčime tie, ktorých p-hodnota [2] prekročí stanovenú hranicu pre ponechanie premennej v modeli. V *NCSS* je k testom významnosti parametrov využívaná Waldova štatistika.

[2] p-hodnota resp. dosiahnutá hladina testu slúži na posúdenie výsledku testu hypotézy. Jedná sa o pravdepodobnosť, že by sme za platnosti hypotézy vypozorovali dáta, ktoré by boli s hypotézou vo väčšom rozpore ako analyzovaný dátový súbor, detailne viď [2]

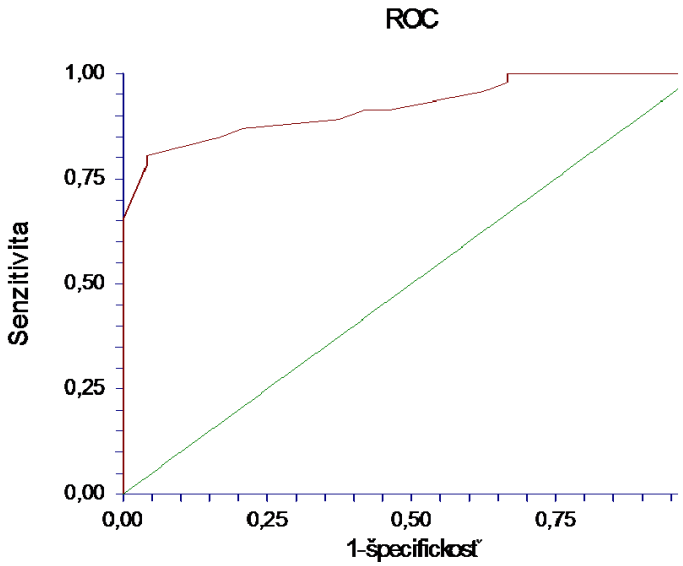
Pri ponechaní všetkých premenných má odhadnutý model tvar

$$18.524 + 4.026 * Adr - 4.678 * BL - 1.2 * DalsieP_1 + 1.201 * DalsieP_2 - 2.409 * DalsieP_3 - 10.375 * Mesto_1 + 0.18 * Mesto_2 - 4.896 * Mesto_3 - 7.576 * Mesto_4 - 3.066 * Mesto_5 - 4.927 * Platenie_1 - 0.612 * Platenie_2 - 2.160 * PC_1 + 1.31 * PC_2 + 8.627 * PC_3 + 4.739 * PC_4 + 0.438 * Pohlavie - 6.447 * Stvrt_1 - 10.567 * Stvrt_2 - 2.085 * SP + 0.641 * TO - 0.108 * vek - 1.200 * Zlavy1 - 2.24 * Zlavy2.$$

Z tohoto modelu sme v prvom iteračnom kroku vylúčili premenné, ktorých p-hodnota prekročila nami zvolenú hladinu $\alpha = 0,05$ na základe Waldovho testu. Ide o premenné $DalsieP_1$, $DalsieP_2$, $Mesto_2$, $Platenie_2$, PC_1, PC_2, $Pohlavie$, To, Vek, $Zlavy1$ a $Zlavy2$. V ďalšej iterácií sme na základe Waldovho testu na rovnakej hladine $\alpha = 0,05$ vylúčili premenné BL, $DalsieP_3$, a SP. Po následnej iterácií už nedošlo k zlepšeniu modelu, preto sme za konečný tvar modelu zvolili nasledujúci:

$$6.84639 + 2.8289 * Adr - 5.03134 * Mesto_1 - 3.01892 * Mesto_3 - 4.29902 * Mesto_4 - 2.11106 * Mesto_5 - 1.23612 * Platenie_1 + 5.2735 * PC_3 + 4.78735 * PC_4 - 5.13568 * Stvrt_1 - 7.860392 * Stvrt_2.$$

V Prílohe D.2 uvádzame výsledné charakteristiky modelu. Všimnime si, že oproti významným premenným v diskriminačnej analýze nám do modelu nevstupujú premenné `TO`, `BL` a ani `Zlavy1` a `Zlavy2`. I napriek tomu sme však pri hraničnej hodnote 0,75 dosiahli úspešnosť klasifikácie pozorovaní približne 85,71%. Za hraničnú hodnotu považujeme pravdepodobnosť, ktorú tolerujeme pre zaradenie pozorovania do skupiny profitabilných zmlúv. Na obrázku 5.1 uvádzame graf Lorenzovej krivky, kde hodnota AUC vyšla 0,92074 a Giniho koeficient rovný 0,84148. Diverzifikačná sila modelu je teda pomerne vysoká.



Obr. 5.1: Graf Lorenzovej krivky

Určený model teraz využijeme na klasifikáciu nových žiadostí, ktoré sme použili v diskriminačnej analýze, viď tabuľka 4.3. Premenné `ADR`, `Mesto`, `PC`, `Stvrt` a `Platenie`, ktoré využívame v modeli logistickej regresie, sme opäť kódovali na dummy premenné. Klasifikáciu - odhad pravdepodobnosti príslušnosti ku kategórii profitabilných zmlúv a indikátor klasifikácie - sme zhrnuli v tabuľke 5.1.

Pozorovanie	**Odhad pravdepodobnosti**	**Klasifikácia na regresie**
71.	0,971959	1
72.	0,999999	1
73.	0,570664	0
74.	0,999945	1
75.	0,942099	1
76.	0,999991	1
77.	0,998218	1
78.	0,001429	0
79.	0,986067	1
80.	0,010387	0
81.	0,936147	1
82.	0,859952	1

Tabuľka 5.1: Klasifikácia nových žiadostí pomocou logistickej regresie

Ako hraničné skóre sme volili hodnotu 0,75 v súlade s aplikáciou pre už zaradené žiadosti. Pozorovania s odhadom pravdepodobnosti vyšším ako táto hodnota by sme zaradili do skupiny profitabilných zmlúv. Pri pozorovaní 79, ktorého klasifikácia bola sporná v diskriminačnej analýze, nám vychádza zaradenie medzi profitabilné zmluvy. Rovnaké zaradenie nám vyšlo v prípade kanonickej diskriminačnej analýzy a naopak medzi neprofitabilné ho zaradila metóda kvadratických diskriminačných funkcií. Ďalším pozorovaním, u ktorého sa líši výsledok klasifikácie s diskriminačnou analýzou, je 82. pozorovanie. Ide o klienta z Prahy, žijúceho v rizikovej oblasti, s nízkou poistnou čiastkou. Dá sa predpokladať, že by klient platil štvrťročne zvýšené poistné, práve za rizikovejšiu oblasť. V prípade poistnej udalosti by však bola vyplácaná nižšia poistná čiastka, mohlo by to teda indikovať ziskovú poistnú zmluvu. Na druhej strane však odhad pravdepodobnosti je nižší ako 90% a v prípade opakovania sa poistných udalostí, ktoré v rizikovej oblasti hrozia vo zvýšenej miere, by sa zmluva mohla taktiež dostať medzi neziskové zmluvy. Priklonili by sme sa teda k zaradeniu medzi neziskové zmluvy, kde nám vyšla zhoda v klasifikácií podľa kanonickej i kvadratickej diskriminačnej analýzy.

Celkovo nám zhodná klasifikácia podľa logistickej regresie, kvadratických diskriminačných funkcií a na základe kanonickej diskriminačnej analýzy, vyšla pre 10 nových žiadostí z 12, čo je približne 83% úspešnosť zhodnej klasifikácie. Výsledky metód môžme teda považovať za obdobné. Logistická regresia kladie na dáta menej striktné požiadavky ako diskriminačná analýza. I napriek tomu sa však voľba vhodnej metódy pre tvorbu prediktívneho modelu môže líšiť, a to najmä v závislosti na skúmaných dátach z rôznych oblastí poistenia.

Kapitola 6

Rola Business Intelligence

6.1 Business Intelligence a jeho štruktúra

Definícia a nastavenie vhodných prediktívnych modelov (alebo procesov využitia expertných pravidiel, prípadne skĺbenie oboch) je nevyhnutnou súčasťou fraud managementu. K efektívnemu fungovaniu však musí byť zakomponovaný do uceleného systému pre monitorovania a hodnotenie rizík podvodov. A ako sa ukazuje v praxi, kľúčová je pritom predovšetkým jeho integrácia do sveta dátových služieb spoločnosti. Dnes sa už na trhu síce vyskytuje množstvo fraud management systémov, avšak poisťovne sa pri ich využívaní často stretávajú so zásadnými problémami, ku ktorým patria:

- Neschopnosť pripraviť vhodné vstupné dáta, či obmedzovanie modelov len na určité segmenty poistenia.
- Obmedzené možnosti správneho vyhodnotenia výsledkov kvantifikácie rizík.
- Chýbajúce prepojenie na ďalšie dátové služby spoločnosti, napríklad reportovanie.
- Neuspokojivá koordinácia s ďalšími zložkami fraud managementu, ako napríklad nevyužívanie výsledkov v oddelení likvidácie poistných udalostí.
- Nedostatočná automatizácia systému.
- Nízka úroveň prispôsobivosti systému na meniace sa podmienky.

Aby k týmto problémom nedochádzalo, je nevyhnutné, aby poisťovne k integrácií týchto systémov pristupovali koncepčné v rámci riadenia celého svojho **Business Intelligence**, ďalej BI.

Podľa [14] BI ako pojem zahŕňa kombináciu produktov, aplikácií, technológií a osvedčených metód na organizáciu kľúčových informácií, ktoré management potrebuje pre optimalizáciu rozhodovania a zlepšenie výkonu spoločnosti. Reálne to znamená, že sa BI stará o riadenie dátových služieb vrátane systémov a aplikácií. Zabezpečuje tiež využitie obsiahnutých informácií v jednotlivých útvaroch spoločnosti. Cieľom BI je najmä dosiahnutie lepšieho usporiadania dát, ktoré sú

obsiahnuté v zdrojových systémoch (napr. účtovnícke systémy), zaistiť ich zjednotenie z viacerých zdrojov a transformovať ich do využiteľnej podoby. Takto reorganizované dáta je možné potom využiť pre strategické rozhodnutia spoločnosti, rôzne analýzy a reportovania. K základným zložkám v rámci BI štruktúry patria:

- *Budovanie a správa dátového skladu* (Data Warehouse - DWH) ako základného dátového úložiska pre analýzu dát.
- *Budovanie a správa dátových tržíšť, datamártov* (Data Mart) ako tematicky orientovaných dátových skladov určených na sprostredkovanie informácií pre konkrétnu službu, či určité oddelenie podniku.
- *Tvorba nadstavbových OLAP aplikácií* (Online Analytical Processing). Jedná sa o technológiu multidimenzionálnych databáz a analytických služieb, ktoré slúžia na analýzu veľkého množstva údajov.
- *Data mining* (dolovanie dát).
- *Prediktívne analýzy* (analýza dát vo viacrozmernej štruktúre).
- *Vytváranie správ a reporting* (prezentácia dát užívateľom).

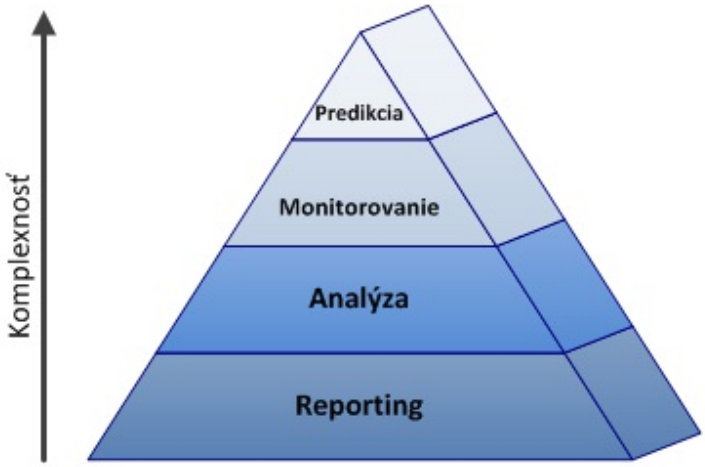
Tieto zložky sa dajú využiť v širokej oblasti procesov poisťovne a podporujú jej každodenný business. Vyplýva z toho tiež, že BI je v podstate prirodzeným spojením sveta byznysu, analýzy, informačných technológií a systémov, kam svojou povahou spadá aj systémove riešenie fraud managementu.

6.2 Systémový koncept fraud managementu

Správny BI systémový koncept fraud managementu, ďalej FMS, by mal podľa nášho názoru spĺňovať niekoľko základných kritérií:

- Musí vedieť využívať dáta zo zdrojových systémov poisťovne a zistené údaje musí vedieť ponúkať užívateľom procesov.
- Musí byť napojený na ďalšie dátové služby a umožňovať flexibilné zmeny, ktoré by mohli byť riadené samotnými užívateľmi.
- Z pohľadu informačného pokrytia, viď obrázok 6.1, musí riešenie dátovo i systémovo pomáhať riešiť otázky vo všetkých základných vrstvách, ktorými sú:
  - Reporting - čo sa stalo? (Aké podvody sa udiali? Ako pracujú sprostredkovatelia a likvidátori?).
  - Analýza - prečo sa to stalo? (Aké sú indikátory podvodov? Ako je možné podvodom zabrániť a ako ich včas odhaliť?).

- Monitorovanie - čo sa deje teraz? (Aké podvody sú zaznamenávané? Aká z nich plynie strata pre poisťovňu? Aká je úspešnosť detekcie, vyšetrovania a ako tieto činnosti prebiehajú?).
- Predikcia - čo sa môže stať (Aké je riziko podvodu na zmluve? Aké straty z podvodu je možné očakávať v budúcnosti?).



Obr. 6.1: Informačné pokrytie. Zdroj: autor

Je teda zjavné, že základným prvkom FMS je okrem výpočtovej aplikácie, kde sa uplatňujú prediktívne modely alebo expertné pravidlá na odhaľovanie podvodov, taktiež podkladová dátová základňa a nadväzujúce služby.

Ideálny koncept FMS, ktoré by poisťovne podľa nás mali naplniť, navrhujeme na obrázku 6.2. Jeho základnými komponentmi by mali byť:

1. **FMS datamárt** ako hlavný zdroj dát (poistné zmluvy, klienti; ich aplikačné a behaviorálne atribúty). Z pohľadu architektúry dátových služieb je datamárt vybudovaný nad dátovým skladom spoločnosti, odkiaľ sa dáta preberajú, transformujú a spracovávajú pomocou ETL[1] nástrojov do už spomínaných dátových skladov. Dátový sklad by mal byť optimalizovaný na čítanie dát, vyhľadávanie, zložité analýzy a taktiež zrozumiteľný pre užívateľov. Na druhú stranu, množstvo poisťovní vlastný dátový sklad ako základné zhromaždisko dát zo zdrojových systémov nemá. V tomto prípade sa samozrejme z finančných dôvodov núka vystavať FMS datamárt priamo nad zdrojovými systémami. Datamárt by sa potom mohol stať prirodzenou dátovou základňou i pre ďalšie účely (finančné a obchodné riadenie,

[1] Jedná sa o skratky slov Extraction, Transformation a Loading. V prvej fáze tohto procesu - extrakcii sú vybrané potrebné dáta zo zdrojových dát. Nasleduje transformácia, počas ktorej dochádza k overovaniu a čisteniu dát (doplnenie chýbajúcich hodnôt, odstránenie preklepov, zjednotenie formátov atď), dátovej konsolidácií a podobne. V poslednej fáze sú dáta premiestňované do dátového skladu.

risk management, marketing a iné). Zdrojom datamártu môžu byť okrem zdrojových systémov (systém pre účtovníctvo, žiadosti, ziskateľský systém, atď.) i ďalšie vstupy:

- Od užívateľov (zahŕňajú rôzne útvary fraud management procesu). Môže ísť o priame vstupy, ku ktorým je možné zaradiť udržovaný black list (zoznam podozrivých klientov), výsledky likvidácie poistných udalostí a jednotlivých šetrení, záznamy pracovníkov call centra alebo iného kontaktu s klientmi. Ďalej sem môžme zaradiť vstupy z analýz väzieb medzi klientmi alebo s poistnými sprostredkovateľmi, vstupy z analýzy sociálnych sietí (pokiaľ je k dispozícii) a taktiež text mining (proces dolovania netriviálnych informácií z textu resp. spracovávanie neštruktúrovaného textu). Vstupom môžu byť tiež pravidlá pre FMS aplikáciu. Máme na mysli namjmä požiadavky na to, čo sa má kontrolovať, aké metódy a s akými parametrami sa majú použiť, aké sú hranice na označenie zmluvy, či poistnej udalosti za podozrivú a mnohé iné.
- Z FMS aplikácie - jedná o zapísanie výsledkov do systému, pretože výstupy z aplikácie by mali byť propagované i útvarom, ktoré s aplikáciou priamo nepracujú.

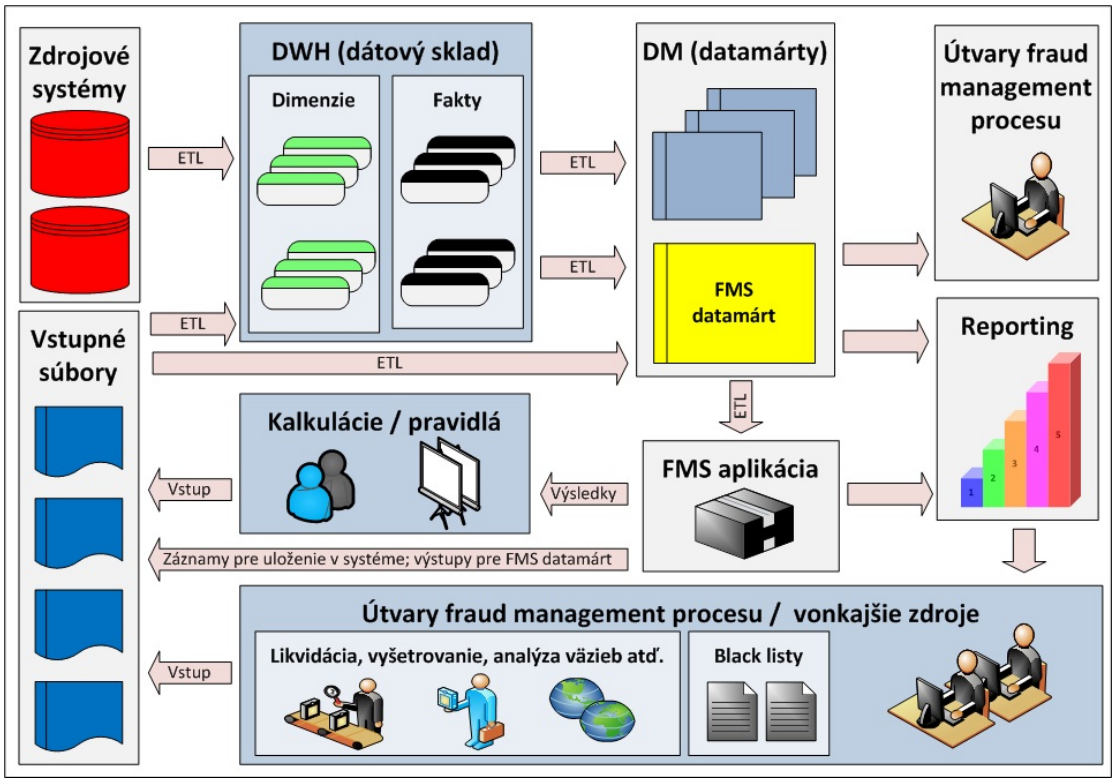
Vo FMS datamárte by malo dochádzať i k naplneniu ďalších BI riešení, akými sú konsolidácia dát a s ňou spojená dátová kvalita. Pokiaľ neexistuje dátový sklad jedná sa napríklad o konsolidáciu klientov. Pokiaľ má poisťovňa viac systémov napr. pre životné a neživotné poistenie, potom môže dôjsť k tomu, že jedna osoba je v týchto systémoch vedená nie ako jeden klient, ale viac klientov s rôznymi produktami. V procese fraud managementu je potom dôležité klientov správne skonsolidovať, t. j. napárovať na jednu osobu. Cieľom je označiť podvodné alebo podozrivé jednanie klienta na jednom produkte a na všetkých jeho zvyšných produktoch v poisťovni. K odberateľom datamártu by mali patriť útvary, služby a aplikácie priamo spojené s fraud managementom využívajúce výsledky k svojím analýzam. Existujú ale aj ďalší možní odberatelia dát ako sú CRM systémy (Customer Relationship Management, respektíve riadenie vzťahov so zákazníkmi), kde sa získané informácie môžu využiť v rámci nových kampaní, v ktorých sa napríklad nebudú oslovovať podozriví klienti.

2. **FMS aplikácia**, kde dochádza k určeniu a selekcii podozrivých zmlúv a poistných udalostí. V rámci aplikácie, ktorej dátovým zdrojom je FMS datamárt, sa využijú navrhnuté modely pomocou štatistických metód a taktiež pravidlá detekcie poistných podvodov pomocou data mingových techník. Takéto aplikácie si môžu poisťovne buď kúpiť, ale potom je potrebné venovať značný priestor integrácii nad FMS datamártom a prispôsobeniu konkrétnym potrebám poisťovne, alebo si ich môžu priamo vyvinúť (napr. pomocou modulov štatistických softvérov).

3. **Reportovanie** ako výstup z uskutočnených analýz vo FMS aplikácii a z FMS datamártu. Výsledky môžu využiť primárni odberatelia, ku kto-

rým patria zástupcovia fraud managementu (tým likvidácie a vyšetrovania, tým taxácie a správy, ALM tým, tým call centra, produktový management) a sekundárni odberatelia ako napríklad manažérsky reporting.

4. **Kalkulácie a nastavenie pravidiel** ako užívateľské vstupy a parametre pre výpočtové modely. V tejto fáze systému je priestor na nastavenie nových pravidiel pre výpočtové modely (napríklad spustiť v aplikácií logistickú regresiu s určitými parametrami pre daný výpis atribútov, či aplikovať model na dáta kapitálového životného poistenia). Užívateľsky riadená by mala byť aj frekvencia spúšťania analýz a určenie odberateľov jednotlivých analýz.



Obr. 6.2: Systémová architektúra FMS. Zdroj: autor

Popísaný koncept splňuje požiadavky na zjednotenie a efektívne využitie dát zo zdrojových systémov. Pomocou FMS aplikácie umožní poisťovniam odhaľovať podozrivé udalosti a v rámci reportovania a kalibrácie ponúka priestor rozšíriť funkčnosť celého procesu na nové druhy podvodov, ktorým je a bude poisťovňa v rozvíjajúcej sa spoločnosti vystavená. Tento koncept teda môže byť svojou štruktúrou podporným systémom pre všetky zložky fraud managementu.

Naplnenie konceptu FMS je samozrejme pre poisťovne jednorázovo finančne nákladnejšie než vybudovať separátne riešenie prostredníctvom zakúpenia samostatnej fraud management aplikácie a zaistenia základného prísunu dát. Ak si však uvedomíme ako rýchlo sa menia podmienky a celkový dátový svet v každej poisťovni, potom je zrejmé, že sa nekoncepčný prístup stáva len krátkodobým riešením. Jeho efektivita bude po nasadení v nasledujúcom období klesať a toto

riešenie sa bez dodatočných nákladov stane neprínosným. K dôvodom prečo by mali poisťovne tento koncept zaviesť do praxe patrí:

- Zvýšenie zisku v zmysle zníženia straty plynúcej z úspešne uskutočnených poistných podvodov.
- Zvýšenie efektivity celého procesu fraud managementu od uzatvárania nových kontraktov, likvidáciu až po vyšetrovanie podvodov.
- Úspora nákladov. Príkladom by mohla byť úspora vďaka nahradeniu práce likvidátorov, ktorí detekujú podozrivé udalosti automatizovaným procesom.

Naviac si myslíme, že ucelený koncept riadenia fraud managementu vystavaný práve na procesoch dátových skladov spojených so službami, ktoré prináša BI, by mal priniesť poisťovniam radu ďalších výhod. Napomôcť likvidátorom a vyšetrovateľom pri prešetrovaní poistných udalostí, či odhadnúť očakávané straty z podvodov. Je však nutné zdôrazniť, že samotné zakúpenie FMS aplikácie na odhalenie podozrivých zmlúv stačiť nebude. Ako sme spomínali v predchádzajúcich kapitolách, rovnako dôležité je i použitie správnych data miningových techník a výpočtových modelov. Takáto kombinácia môže byť pre každú poisťovňu tým najlepším predpokladom k rastúcej úspešnosti v boji s poistným podvodom.

Záver

V tejto práci sme sa venovali problematike poistných podvodov a možnostiam ako efektívne riadiť fraud management v poisťovni. Cieľom práce bolo nájsť spôsoby tvorby prediktívnych modelov s využitím mnohorozmerných metód štatistickej analýzy a ukázať možnosť ich integrácie do dátových systémov spoločnosti pomocou Business Intelligence.

Na úvod sme sa snažili nájsť odpovede na otázky: Prečo ľudia páchajú poistné podvody? Aké typy poistných podvodov sa vyskytujú v našej spoločnosti, či ako sa poistitelia snažia odhaľovať tieto podvody? Zistili sme, že hodnoty odhalených poistných podvodov z roka na rok rastú, čo značí, že poistitelia túto problematiku neberú na ľahkú váhu. Práve naopak, pomocou dátovej analýzy a procesných nastavení sa snažia detekovať podozrivé udalosti a zabraňovať tým neoprávneným výplatám poistného plnenia. My sme sa zamerali na oblasť odhaľovania podvodných jednaní pomocou konštrukcie prediktívnych modelov na základe klasifikačných metód a dataminigových techník.

Druhá kapitola bola zameraná na priblíženie praktickej úlohy z vybraného typu poistenia - poistenia domácnosti. Úlohu sme zostavili prepojením predikcie pravdepodobnosti poistného podvodu s predikciou profitabilného správania sa zmlúv v dlhodobom horizonte na základe ich škodného pomeru. Našim zámerom bolo nájsť také modely, ktoré budú na základe odhadu budúceho vývoja pomáhať pri schvaľovaní žiadostí o poistenie domácnosti. A to takým spôsobom, aby do portfólia neboli zaradené výrazne neziskové zmluvy alebo zmluvy s vysokou pravdepodobnosťou možného poistného podvodu. Priblížili sme dátový súbor, ktorý nám ďalej slúžil ako historická dátova základňa, na základe ktorej sme navrhovali modely klasifikácie nových žiadostí.

Vhodná selekcia premenných pred samotnou konštrukciou prediktívneho modelu, často vedie k zjednodušeniu výpočtu a interpretácií navrhnutého modelu. Preto sme sa z metód mnohorozmernej štatistiky zamerali v tretej kapitole najprv na analýzu hlavných komponentov. Pomocou tejto metódy sa nám podarilo znížiť počet meraných znakov v našej úlohe na jednotlivých žiadostiach o poistenie domácnosti, z pôvodných 18 pozorovaných znakov historickej dátovej základne na 15. Využili sme pri tom maticu koeficientov hlavných komponentov, ktorú sme spočítali pomocou softvéru *Mathematica 8.0*. Vytvorili sme všeobecný kód umožňujúci rozšírenie úlohy aj na iné dátové matice a typ poistenia.

V ďalšej časti sme pristúpili k samotnej klasifikácií nových žiadostí. Využili sme k tomu kanonickú diskriminačnú analýzu a kvadratické diskriminačné funkcie. Predstavili sme tiež primárnu úlohu diskriminačnej analýzy, tak ako ju v 30. rokoch 20. storočia definoval Ronald Aylmer Fisher, teda schopnosť meraných

znakov odlíšiť jednotlivé klasifikačné skupiny. Dospeli sme k tomu, že na odlíšenie nami stanovených skupín, nám postačí 9 meraných znakov. Tieto znaky sme ďalej využili pri klasifikácií nových pozorovaní do 2, predom definovaných, skupín. Obe metódy diskriminačnej analýzy dosiahli úspešnosť klasifikácie historických pozorovaní vyššiu ako 80%. Vybranú skupinu nových žiadostí klasifikovali obe metódy rovnakým spôsobom vo viac než 90% prípadov. Metódy diskriminačnej analýzy sme, rovnako ako pomocné výpočty, aplikovali zostavením kódu v softvéri *Mathematica 8.0*.

Ako alternatívu k metódam diskriminačnej analýzy sme v piatej kapitole predstavili model logistickej regresie, ktorého výhodou je, že nie je obmedzený tvarom rozdelenia. Konštrukciu modelu sme previedli metódou stepwise v štatistickom softvéri *NCSS 2007*. Kvôli kategoriálnemu charakteru vstupných premenných sme ich najprv prekódovali na dummy premenné. Úspešnosť klasifikácie historických pozorovaní taktiež dosiahla hodnoty vyššej ako 80% pre stanovenú hodnotu pravdepodobnosti. Pri klasifikácií nových žiadostí sme dosiahli klasifikačnú zhodu s metódami diskriminačnej analýzy v 85% prípadov, pričom ale náš model logistickej regresie využíva na klasifikáciu menej pozorovaných znakov. Pre nami vybraný typ poistenie sú teda obe metódy takmer zhodné a tým pádom predstavujú hodnotné alternatívy k praktickej integrácii podľa voľby poisťovne.

Na záver práce sme sa venovali predstaveniu pojmu Business Intelligence, pomocou ktorého je možné nájdené prediktívne modely zahrnúť do celistvého systému monitorovania poistných podvodov a dátových služieb poisťovní. Navrhli sme systémový koncept fraud managementu, ktorý by mal zahrňovať efektívne využívanie dát zo zdrojových systémov poisťovne, mal by byť flexibilný a dostupný užívateľom v rámci poisťovne. Dôležité je teda prepojenie výpočtových aplikácií, ku ktorým patria i prediktívne modely, s dostatočnou dátovou základňou a nadväzujúcimi službami, ktoré sú nevyhnutné pre efektívne fungovanie poisťovne.

Vzhľadom k tomu, že jednotlivé oblasti poistenia majú svoje špecifiká a ľudia resp. organizované skupiny sú pri páchaní poistných podvodov čoraz vynaliezavejší, je v problematike poistných podvodov stále priestor na zefektívňovanie fraud managementu. Myslíme si, že tou správnou cestou je popri integrácii nových procesných nastavení najmä rozvoj v oblastiach aplikácie dataminigových techník. Veríme, že zostavením prediktívnych modelov a navrhnutím fraud management systému bol dosiahnutý nielen cieľ našej práce, ale predovšetkým, že sme našou prácou pozitívnym spôsobom prispeli v nekonečnom boji s poistnými podvodmi.

Literatúra

[1] Anděl, J.: Matematická statistika, Nakladatelství technické literatuty SNTL, 1985.

[2] Anděl, J.: Základy matematické statistiky, MATFYZPRESS, 2007. ISBN 80-7378-001-1.

[3] Česká asociace pojišťoven. http://www.cap.cz/ItemP.aspx?t=0

[4] Härdle, W., Simar, L.: Applied Multivariate Statistical Analysis, 2003. http://www.stat.wvu.edu/~jharner/courses/stat541/mva.pdf

[5] Hebák, P.: Vícerozměrné statistické metody (1). Informatorium, Praha, 2007. ISBN 978-80-7333-056-9.

[6] Hosmer, D. W., Lemeshow, S.: Applied Logistic Regression, 2nd ed., John Wiley & Sons, Inc., 2000.

[7] Húsek, D.,Řeznáková, H., Snášel, V.: Shluková analýza dát, Professional Publishing, 2007. ISBN 978-80-86946-26-9.

[8] Insurance Europe. http://www.insuranceeurope.eu/

[9] Pojištění domácnosti a nemovitosti. http://www.mesec.cz/bydleni/pojisteni-domacnosti-a-nemovitosti/pruvodce/

[10] Program NCSS 2007 - Help.

[11] Quantitative Impact Study. EIOPA, 2011.http://eiopa.europa.eu/fileadmin/tx_dam/files/publications/reports/QIS5_Report_Final.pdf

[12] Trestní zákoník 2012 (zákon č. 40/2009 Sb.).

[13] Wilhelm, W. K.: The Fraud Management Lifecycle Theory. Journal of Economic Crime Management, 2004. https://www.utica.edu/academic/institutes/ecii/publications/articles/BA309CD2-01B6-DA6B-5F1DD7850BF6EE22.pdf

[14] Williams, N., Williams. S.: The profit Impact of Business Intelligence, Morgan Kaufmann, 2006.

[15] Wolfram Research, Inc. (2010): Wolfram Mathematica, verze 8.0, Champaign.

Zoznam tabuliek

Zoznam obrázkov

Prílohy

Príloha A.1: Dátová matica

Pozorovanie	Profit	Fraud	Stvrt	Mesto	TO	PC	Titul	Vek	Pohlavie	Adr	SP	Bydlisko	Platenie	BL	DalsieP	Cudzinec	Zlavy1	Zlavy2
1.	1	1	1	1	1	5	1	39	1	1	2	1	1	2	2	2	2	2
2.	1	1	1	1	1	5	1	52	1	1	2	1	1	2	1	2	2	1
3.	1	1	1	1	1	5	1	70	2	1	2	1	2	2	1	2	2	1
4.	1	1	1	1	2	5	1	43	1	2	2	1	1	2	1	2	1	2
5.	1	1	1	1	1	4	2	29	1	2	2	1	1	2	3	1	1	1
6.	1	1	2	1	1	3	2	61	1	2	1	1	1	2	2	2	2	1
7.	1	1	2	1	1	4	1	40	2	1	2	1	1	2	2	2	2	1
8.	1	1	2	1	2	3	2	43	1	2	1	2	1	2	1	2	1	1
9.	1	1	2	1	2	3	2	35	2	2	2	1	1	1	4	2	2	1
10.	1	1	3	1	2	2	1	26	1	2	2	1	1	2	1	2	1	1
11.	1	1	1	2	1	5	2	43	2	1	2	1	1	2	1	1	2	2
12.	1	1	2	2	1	4	1	36	1	2	2	1	1	2	2	2	1	1
13.	1	1	2	2	1	3	1	25	2	1	2	1	1	1	3	2	2	2
14.	1	1	2	1	2	5	2	50	1	2	2	1	1	2	1	2	2	2
15.	1	1	2	3	1	5	1	38	1	1	1	1	1	2	2	2	2	2
16.	1	1	2	3	1	4	1	28	1	1	1	1	1	2	2	2	1	2
17.	1	1	2	3	1	3	1	35	2	2	2	2	2	2	2	2	1	1
18.	1	1	2	3	1	3	1	55	2	1	2	1	2	2	3	2	2	2
19.	1	1	3	3	1	2	2	63	1	1	2	1	1	2	4	2	2	2
20.	1	1	1	4	1	5	1	40	1	1	2	1	1	2	2	2	1	1
21.	1	1	1	4	1	4	1	34	1	1	1	1	1	2	1	2	2	1
22.	1	1	1	4	1	3	2	44	1	1	2	1	1	2	1	2	2	2
23.	1	1	2	4	1	3	2	33	2	1	2	1	1	2	2	2	1	2
24.	1	1	2	4	1	3	2	55	1	1	2	1	3	2	4	2	2	2
25.	1	1	1	5	1	5	1	72	1	1	2	1	2	2	1	2	2	1
26.	1	1	1	5	1	5	2	34	1	1	2	1	1	2	2	2	2	2
27.	1	1	1	3	1	4	2	53	2	1	2	1	1	2	2	2	2	2
28.	1	1	2	5	1	2	2	27	2	1	2	1	1	2	1	2	1	1
29.	1	1	1	1	1	5	2	44	1	1	2	1	1	2	2	2	2	2
30.	1	1	2	6	1	3	2	43	2	1	2	1	2	2	2	2	1	2
31.	1	1	3	6	1	2	2	32	2	1	2	1	1	2	1	2	2	1
32.	1	1	3	6	1	2	2	26	1	1	2	1	2	2	2	2	2	2
33.	2	1	1	1	1	4	1	32	2	1	2	1	1	2	2	2	2	2
34.	2	1	2	1	1	4	1	35	1	1	2	1	2	2	1	2	2	1
35.	2	1	3	1	1	3	2	37	1	2	1	1	3	2	1	2	1	1
36.	2	1	3	1	1	3	2	28	1	1	2	1	1	1	2	2	1	1
37.	2	1	2	3	1	4	2	47	1	2	2	1	1	2	1	2	1	1
38.	2	1	3	3	1	1	2	54	1	1	2	1	2	2	1	2	2	1
39.	2	1	1	4	1	4	1	69	2	1	2	1	2	2	2	2	2	2
40.	2	1	1	4	1	3	2	35	2	1	2	1	1	1	2	2	1	1
41.	2	1	1	4	1	4	2	32	1	1	1	1	1	1	3	2	2	2
42.	2	1	3	4	1	2	2	57	1	1	2	1	1	2	1	2	2	2
43.	2	1	1	5	1	4	1	39	1	1	1	1	1	2	1	1	1	2
44.	2	1	2	5	1	4	2	46	1	2	2	1	2	2	3	2	1	1
45.	2	1	2	6	1	4	2	62	2	1	1	1	1	2	1	2	2	1
46.	2	1	2	6	1	3	1	30	1	1	2	1	1	2	2	2	2	1
47.	1	2	3	1	1	2	2	47	2	1	2	1	2	2	1	1	2	2
48.	1	2	3	2	2	4	2	41	2	2	2	1	1	2	1	2	1	1
49.	2	2	3	1	1	3	1	32	1	1	2	1	2	2	1	2	2	1
50.	2	2	3	3	1	3	2	27	1	2	2	1	3	1	4	2	1	1
51.	3	1	2	2	1	3	2	58	2	1	2	1	3	2	2	2	2	1
52.	3	1	3	3	2	1	1	36	1	2	1	1	1	1	2	2	1	1
53.	3	1	3	3	1	2	2	38	1	2	2	1	1	2	2	2	1	2
54.	3	1	2	4	2	2	1	24	1	2	2	1	1	2	2	2	1	1
55.	3	1	2	4	1	3	1	41	2	1	1	1	1	1	4	2	2	2
56.	3	1	3	4	1	2	1	59	1	2	2	1	3	2	1	2	1	2
57.	3	1	2	5	1	2	2	28	1	1	2	2	1	2	2	2	1	1
58.	3	1	2	5	1	2	2	38	1	2	1	1	3	2	4	2	1	1
59.	3	1	3	6	2	1	1	37	1	2	2	1	1	2	3	2	1	1
60.	3	2	3	3	1	3	2	30	1	2	1	1	1	1	1	2	2	1
61.	3	2	2	4	1	1	1	38	1	2	2	1	2	1	2	2	1	1
62.	4	1	2	2	2	2	2	49	1	2	2	1	2	2	2	2	1	1
63.	4	1	3	2	1	3	2	34	1	1	1	1	3	2	1	2	1	1
64.	4	1	3	3	1	2	2	40	2	1	2	1	2	2	2	2	1	1
65.	4	1	3	4	2	2	2	39	2	2	2	1	2	2	1	2	1	1
66.	4	1	2	6	1	2	2	53	2	2	2	1	1	2	1	2	1	1
67.	4	1	3	6	1	1	2	50	1	2	2	1	2	2	1	2	1	1
68.	4	1	3	6	1	1	2	23	1	1	2	1	1	1	4	1	2	1
69.	4	2	2	6	2	2	2	47	1	2	2	2	3	2	1	2	2	1
70.	4	2	3	6	1	2	2	29	1	1	2	1	1	1	4	2	1	1

Príloha B.1: Analýza hlavných komponentov (kód)

Kompletný nižšie uvedený kód je k dispozícii taktiež na priloženom CD, v súbore *Analýza_ hlavných_ komponent.nb*.

Načítanie dát:

```
SetDirectory[NotebookDirectory[]];
data = Import["Data.xlsx"][[1]];
d = Take[data, 71];
nazvy = First[data];
datovamatica = Rest[d];
tdatovamatica = Transpose[datovamatica];
pocetpozorovani = Dimensions[datovamatica][[1]]
pocetznakov = Dimensions[datovamatica][[2]]
```

Výberový priemer a smerodajná odchýlka:

```
Mean[datovamatica];
StandardDeviation[datovamatica];
```

Štandardizácia dát:

```
standardizovane = Table[(datovamatica[[i]] - Mean[datovamatica])/
   StandardDeviation[datovamatica], {i, 1, pocetpozorovani}];
```

Kovariančná matica:

```
Needs["MultivariateStatistics‘"]
kovariancnaM = Covariance[standardizovane];
TableForm[Round[kovariancnaM, 0.01],TableHeadings ->{nazvy,nazvy}]
```

(výstupný formát)

Spektrálny rozklad:

```
kovariancna = Covariance[standardizovane];
system = Eigensystem[kovariancna]; (vlastné čísla a vlastné vektory)
lambda = DiagonalMatrix[system[[1]]]; (diagonálna matica)
```

Akú časť variability vysvetľujú jednodlivé vlastné čísla:

```
vlcisla = system[[1]];
percenta = Table[vlcisla[[i]]/Total[vlcisla]*100,{i,1,pocetznakov}];
kumulativne = Accumulate[percenta];
cislo = Table[i, {i, 1, pocetznakov}];
tabulka = Transpose[{vlcisla, percenta, kumulativne}];
TableForm[tabulka,TableHeadings->{cislo,{"vlastné číslo","percentá",
  "kumulatívne"}}] (výstupný formát)

qm = ListPlot[kumulativne, DataRange -> {0, pocetznakov},
  AxesLabel->{Style["m", Large], Style["q(m)", Medium]},
       PlotStyle->{Darker[Green],PointSize[Medium]},Filling->Axis,
```

```
        LabelStyle->Directive[Medium]] (graf funkcie q(m))

vl = Transpose[{Range[pocetznakov], vlcisla}];
vl1 = ListPlot[vl, DataRange->{0, pocetznakov},
  AxesOrigin->{-1, 0}; AxesLabel -> {"vl.c.","Hodnota"},
       Filling->Axis, PlotStyle -> PointSize[Medium],
            LabelStyle->Directive[Medium]] (scree graf)
```

Koeficienty, vlastné vektory:

```
p = Transpose[system[[2]]]; (vlastné vektory)
komponenty1=Table[Subscript[Z,i],{i,1,pocetznakov}];koeficienty = p;
TableForm[Round[koeficienty,0.001],TableHeadings->{nazvy,komponenty1}]
(výstupný formát)
```

Hlavné komponenty a ich rozptyly:

```
hk = Transpose[standardizovane.p]; (hlavné komponenty)
rozptyly = Table[Variance[hk[[i]]], {i,1,pocetznakov}]
stdata = Transpose[standardizovane];
rozptylydat = Table[Variance[stdata[[i]]], {i,1,pocetznakov}];
Total[rozptyly]==Total[vlcisla]==pocetznakov==Total[rozptylydat]
(porovnanie celkovej variability)
```

Príloha C.1: Mnohorozmerná normalita (kód)

Kompletný nižšie uvedený kód je k dispozícii taktiež na priloženom CD, v súbore *Test_ normality.nb*.

Načítanie dát, vstupné parametre:

```
SetDirectory[NotebookDirectory[]];
Needs["MultivariateStatistics`"]
data = Import["data.xlsx"][[2]];
d = Take[data, 71];
datMatica = Rest[d];
α = 0.05;
n = Dimensions[datMatica][[1]];
p = Dimensions[datMatica][[2]];
priemer = Mean[datMatica];
s = Covariance[datMatica]; (výberová kovariančná matica)
sn = Sum[Transpose[{datMatica[[i]] - priemer}].
  {(datMatica[[i]]-priemer)},{i,1,n}]/(n-1);
  (alternatívny výpočet kovariančnej matice)
```

Koeficient šikmosti:

```
b1 = (1/n^2)*Sum[Sum[({datMatica[[i]] - priemer}.Inverse[sn].
  Transpose[{datMatica[[j]] - priemer}])^3,
       {j,1,n}], {i,1,n}]//First //First
bb1 = MultivariateSkewness[datMatica]
```

Porovnanie so zabudovanou funkciou na výpočet mnohorozmernej šikmosti:

```
bb1 == b1

statistikaV = (n*b1)/6
stvolnosti = p*(p + 1)*(p + 2)/6 (počet stupňov voľnosti)
k1 = InverseCDF[ChiSquareDistribution[stvolnosti], 1 - α /2]
(kvantil chí-kvadrát rozdelenie)
statistikaV > k1 (test)
```

Koeficient špicatosti:

```
b2 = (1/n)*Sum[({datMatica[[i]] - priemer}.Inverse[sn].
   Transpose[{datMatica[[i]] - priemer}])^2,
         {i, 1, n}]//First//First
bb2 = MultivariateKurtosis[datMatica]
```

Porovnanie so zabudovanou funkciou na výpočet mnohorozmernej špicatosti:

```
bb2 == b2

statistikaU = Sqrt[n/(8*p*(p + 2))]*(b2 - p*(p + 2))
k2 = InverseCDF[NormalDistribution[], 1 - α /2]
(kvantil normálneho rozdelenia)
statistikaU > k2 (test)
```

Príloha C.2: Diskriminačná analýza (kód)

Kompletný nižšie uvedený kód je k dispozícii taktiež na priloženom CD, v súboroch *Diskriminačná_ analýza.nb* a *Diskriminačná_ analýza-vylúčené_ premenné.nb*.

Načítanie dát, vstupné parametre:

```
SetDirectory[NotebookDirectory[]];
pocetskupin = 2; pocetznakov = 13;
```

(resp. počet znakov 9, v prípade DA s vylúčenými premennými)

```
nacitanie = Import["data.xlsx"][[3]];
```

(resp. `[[4]]` v prípade diskriminačnej analýzy s vylúčenými premennými)

```
bezhlavicky = IntegerPart[Rest[nacitanie]];
profit = bezhlavicky[[All, 1]];
skupiny = bezhlavicky[[1 ;; 70, 2 ;; pocetznakov + 1]];
```

(dátova matica bez premennej `ProfitF`)

```
pocty = Table[Count[profit,i], {i, pocetskupin- 1,0,-1}]
```

(počty objektov v skupinách)

Celkový priemer a priemery znakov v skupinách:

```
celkovyp=Table[N[Mean[skupiny[[All,i]]]],{i,1,pocetznakov}]
priemer={} ;a = 1; b = 0; p = 0;
While[p < pocetskupin, { p++, b = b + pocty[[p]],
   Print[AppendTo[priemer,
         Table[N[Sum[skupiny[[j,l]], {j,a,b}]/pocty[[p]]],
               {l,1,pocetznakov}]]], a = 1 + pocty[[p]]}];
```

Výberové kovariančné matice skupín:

```
kovariancia = {};a = 1; b = 0; p = 0;
While[p < pocetskupin,{p++,b = b + pocty[[p]],
  Print[AppendTo[kovariancia,
    Sum[1/(pocty[[p]]-1)*(Transpose[{skupiny[[j]]-priemer[[p]]}]).
      ({skupiny[[j]]-priemer[[p]]}),{j,a,b}]]],
        a = 1 + pocty[[p]]}];
```

Matica B:

```
maticaB = Sum[pocty[[i]]*(Transpose[{priemer[[i]] - celkovyp}]).
  ({priemer[[i]]-celkovyp}),{i,1,pocetskupin}]//Rationalize;;
```

Matica E:

```
maticaE = Sum[(pocty[[i]]-1)*kovariancia[[i]], {i,1,pocetskupin}];
```

Matica H:

```
H = DiagonalMatrix[Table[Sqrt[maticaE[[i,i]]], {i,1,pocetznakov}]];
```

Riešenie sústavy s normovacou podmienkou:

```
vlcisla = Sort[Eigenvalues[maticaB.Inverse[maticaE]], Greater];
Clear[u,x]; u = Map[x, Range[pocetznakov]];
prava = Table[0, {i,1,pocetznakov}];
```

```
normovanie = (1/(pocetobjektov-pocetskupin))*u.maticaE//Rationalize;
NSolve[Rationalize[(maticaB-
  Eigenvalues[maticaB.Inverse[maticaE]][[1]]*maticaE).u] ==
    prava && normovanie.u == 1, u]
riesenie = u /. %
prve = riesenie[[1]]
```

Stanovenie koeficientov:

```
v1 = -prve.celkovyp
priemerydiskr = Table[v1 + prve.priemer[[i]], i, 1, pocetskupin]
```

(priemery diskriminantov)

```
normovane = (1/Sqrt[(pocetobjektov - pocetskupin)])*H.prve//N
```

(normované koeficienty)

```
kk=(1/Sqrt[(pocetobjektov-pocetskupin)])*Inverse[H].maticaE.prve//N
```

(korelačné koeficienty)

Klasifikácia pozorovaní:

```
kanskory = Table[((v1 + prve.skupiny[[i]])-priemerydiskr[[j]])^2,
   {i,1,pocetobjektov}, {j, 1, pocetskupin}];
priemerydiskr = Table[v1 + prve.priemer[[i]], i, 1, pocetskupin]
```

(matica vzdialeností)

```
chybne=Select[Table[If[Position[kanskory[[i]],Min[kanskory[[i]]]][[1,1]]
  ≠skupina[[i]], i, 0],{i,1,pocetobjektov}],# ≠0 &]
percento = 100*(pocetobjektov - Length[chybne])/pocetobjektov//N
```

(úspešnosť kanonickej DA)

```
determinanty = Table[Det[kovariancia[[i]]],{i,1,pocetskupin}]
inverzne = Table[Inverse[kovariancia[[i]]],{i,1,pocetskupin}];
kvadraticka=Table[(-1/2)*Log[determinanty[[j]]]-
  (1/2)*(skupiny[[i]] - priemer[[j]]).
    inverzne[[j]].(skupiny[[i]]-priemer[[j]]) +
      Log[pocty[[j]]/pocetobjektov],{i,1,pocetobjektov},
        {j,1,pocetskupin}];
```

(kvadratická DA)

```
chybneK= elect[Table[If[Position[kvadraticka[[i]],
  Max[kvadraticka[[i]]]][[1, 1]]≠
    skupina[[i]], i, 0], {i,1,pocetobjektov}],# ≠ 0&]
percentok = 100*(pocetobjektov - Length[chybneK])/pocetobjektov//N
```

(úspešnosť kvadratickej DA)

Klasifikácia nových pozorovaní:

```
nacitanien = First[Import["Datan_da.xlsx"]];
bezhlavickyn = IntegerPart[Rest[nacitanien]];
kvadraticken=Table[-1/2*Log[determinanty[[j]]]-(1/2)*(bezhlavickyn[[i]]-
  priemer[[j]]).inverzne[[j]].(bezhlavickyn[[i]] - priemer[[j]])+
    Log[pocty[[j]]/pocetobjektov], {i,1,Length[bezhlavickyn]},
```

```
      {j,1,pocetskupin}]
Table[If[Position[kvadraticken[[i]],Max[kvadraticken[[i]]][[1,1]]==
   1,1,0], {i, 1, Length[bezhlavickyn]}]
```
(zaradenie do skupín pomocou kvadratickej DA)

```
kanskoryn=Table[((v1+prve.bezhlavickyn[[i]])-priemerydiskr[[j]])^2,
  {i,1,Length[bezhlavickyn]},{j,1,pocetskupin}]
Table[If[Position[kanskoryn[[i]],Min[kanskoryn[[i]]][[1,1]]==
   1,1,0], {i,1,Length[bezhlavickyn]}]
```
(zaradenie do skupín pomocou kanonickej DA)

Príloha C.3: Výberové kovariančné matice skupín

Výberová kovariančná matica, kde `ProfitF`$= 1$.

$$\begin{pmatrix}
0,5295 & 0,0841 & 0,0242 & -0,5758 & -1,2638 & -0,0338 & 0,058 & 0,001 & 0,0899 & -0,0019 & 0,028 & -0,0377 & -0,0715 \\
0,0841 & 3,1329 & -0,2198 & -0,4754 & 0,5425 & 0,0966 & -0,3275 & -0,0266 & 0,1179 & 0,0643 & -0,0039 & 0,014 & 0,0995 \\
0,0242 & -0,2198 & 0,099 & -0,001 & -0,3348 & -0,0164 & 0,0821 & -0,0005 & -0,0338 & -0,0101 & -0,0251 & -0,0256 & -0,0087 \\
-0,5758 & -0,4754 & -0,001 & 1,1324 & 2,343 & -0,0609 & -0,0068 & -0,0338 & -0,1005 & 0,0454 & -0,114 & 0,0522 & 0,0802 \\
-1,2638 & 0,5425 & -0,3348 & 2,343 & 159,803 & 0,2309 & -0,4657 & 0,1715 & 2,1159 & 1,2681 & -0,771 & 2,356 & 0,2647 \\
-0,0338 & 0,0966 & -0,0164 & -0,0609 & 0,2309 & 0,2319 & -0,0483 & 0,0473 & 0,0029 & -0,028 & 0,0396 & 0,0203 & 0,0077 \\
0,058 & -0,3275 & 0,0821 & -0,0068 & -0,4657 & -0,0483 & 0,1971 & -0,0145 & 0,0077 & 0,0068 & 0,0019 & -0,1014 & -0,0831 \\
0,001 & -0,0266 & -0,0005 & -0,0338 & 0,1715 & 0,0473 & -0,0145 & 0,1609 & 0,0164 & 0,0005 & 0,0541 & 0,015 & 0,0068 \\
0,0899 & 0,1179 & -0,0338 & -0,1005 & 2,1159 & 0,0029 & 0,0077 & 0,0164 & 0,3053 & 0,0338 & 0,0541 & 0,0039 & -0,0155 \\
-0,0019 & 0,0643 & -0,0101 & 0,0454 & 1,2681 & -0,028 & 0,0068 & 0,0005 & 0,0338 & 0,099 & -0,1082 & 0,0034 & 0,0087 \\
0,028 & -0,0039 & -0,0251 & -0,114 & -0,771 & 0,0396 & 0,0019 & 0,0541 & 0,0541 & -0,1082 & 0,7691 & 0,0454 & 0,1072 \\
-0,0377 & 0,014 & -0,0256 & 0,0522 & 2,356 & 0,0203 & -0,1014 & 0,015 & 0,0039 & 0,0034 & 0,0454 & 0,2382 & 0,0696 \\
-0,0715 & 0,0995 & -0,0087 & 0,0802 & 0,2647 & 0,0077 & -0,0831 & 0,0068 & -0,0155 & 0,0087 & 0,1072 & 0,0696 & 0,2551
\end{pmatrix}$$

Výberová kovariančná matica, kde `ProfitF`$= 0$.

$$\begin{pmatrix}
0,2446 & -0,1685 & -0,0163 & 0,0054 & -1,0543 & -0,0163 & -0,0163 & 0,0054 & -0,038 & -0,0272 & -0,0707 & -0,0163 & 0,0217 \\
-0,1685 & 2,7808 & 0,0199 & -0,7554 & -1,4167 & -0,1975 & 0,1793 & 0,0851 & -0,3062 & -0,1069 & 0,6612 & -0,154 & -0,1377 \\
-0,0163 & 0,0199 & 0,2156 & -0,038 & -0,0254 & -0,0018 & 0,1141 & 0,0199 & -0,067 & 0,0453 & -0,0996 & -0,0453 & -0,0507 \\
0,0054 & -0,7554 & -0,038 & 0,6359 & 0,1196 & 0,1359 & -0,0815 & -0,0598 & 0,1141 & 0,038 & -0,1359 & 0,0924 & 0,0217 \\
-1,0543 & -1,4167 & -0,0254 & 0,1196 & 101,384 & 1,9746 & 0,8587 & 0,7138 & 3,4094 & 2,1558 & -5,1341 & 0,192 & 1,2464 \\
-0,0163 & -0,1975 & -0,0018 & 0,1359 & 1,9746 & 0,2156 & -0,0598 & 0,0199 & -0,0236 & 0,0453 & -0,0996 & 0,0417 & 0,0362 \\
-0,0163 & 0,1793 & 0,1141 & -0,0815 & 0,8587 & -0,0598 & 0,2446 & 0,0054 & 0,0054 & 0,0163 & -0,1141 & -0,1033 & -0,0217 \\
0,0054 & 0,0851 & 0,0199 & -0,0598 & 0,7138 & 0,0199 & 0,0054 & 0,1721 & -0,0018 & 0,067 & -0,0779 & -0,0236 & -0,0072 \\
-0,038 & -0,3062 & -0,067 & 0,1141 & 3,4094 & -0,0236 & 0,0054 & -0,0018 & 0,6938 & 0,1105 & -0,1214 & 0,0199 & -0,0072 \\
-0,0272 & -0,1069 & 0,0453 & 0,038 & 2,1558 & 0,0453 & 0,0163 & 0,067 & 0,1105 & 0,2156 & -0,2917 & -0,0417 & 0,0072 \\
-0,0707 & 0,6612 & -0,0996 & -0,1359 & -5,1341 & -0,0996 & -0,1141 & -0,0779 & -0,1214 & -0,2917 & 1,346 & -0,0127 & -0,0072 \\
-0,0163 & -0,154 & -0,0453 & 0,0924 & 0,192 & 0,0417 & -0,1033 & -0,0236 & 0,0199 & -0,0417 & -0,0127 & 0,2156 & 0,0362 \\
0,0217 & -0,1377 & -0,0507 & 0,0217 & 1,2464 & 0,0362 & -0,0217 & -0,0072 & -0,0072 & 0,0072 & -0,0072 & 0,0362 & 0,1449
\end{pmatrix}$$

Príloha C.4: Matice E, B a H

Matica $\boldsymbol{B}$:

$$\begin{pmatrix}
11,1918 & 10,8067 & 2,4309 & -19,712 & -44,2376 & -0,746118 & 4,83773 & -0,168478 & 6,47438 & -2,4309 & 2,86413 & -4,50078 & -4,13975 \\
10,8067 & 10,4348 & 2,34726 & -19,0337 & -42,7154 & -0,720445 & 4,67127 & -0,162681 & 6,2516 & -2,34726 & 2,76558 & -4,34591 & -3,99731 \\
2,4309 & 2,34726 & 0,528002 & -4,28152 & -9,60859 & -0,16206 & 1,05078 & -0,0365942 & 1,40626 & -0,528002 & 0,622101 & -0,977588 & -0,899172 \\
-19,712 & -19,0337 & -4,28152 & 34,7185 & 77,9152 & 1,31413 & -8,52065 & 0,296739 & -11,4033 & 4,28152 & -5,04457 & 7,92717 & 7,2913 \\
-44,2376 & -42,7154 & -9,60859 & 77,9152 & 174,857 & 2,94917 & -19,122 & 0,665942 & -25,5912 & 9,60859 & -11,321 & 17,7902 & 16,3631 \\
-0,746118 & -0,720445 & -0,16206 & 1,31413 & 2,94917 & 0,0497412 & -0,322516 & 0,0112319 & -0,431625 & 0,16206 & -0,190942 & 0,300052 & 0,275983 \\
4,83773 & 4,67127 & 1,05078 & -8,52065 & -19,122 & -0,322516 & 2,09115 & -0,0728261 & 2,7986 & -1,05078 & 1,23804 & -1,9455 & -1,78944 \\
-0,168478 & -0,162681 & -0,0365942 & 0,296739 & 0,665942 & 0,0112319 & -0,0728261 & 0,00253623 & -0,0974638 & 0,0365942 & -0,0431159 & 0,0677536 & 0,0623188 \\
6,47438 & 6,2516 & 1,40626 & -11,4033 & -25,5912 & -0,431625 & 2,7986 & -0,0974638 & 3,74539 & -1,40626 & 1,65688 & -2,60367 & -2,39482 \\
-2,4309 & -2,34726 & -0,528002 & 4,28152 & 9,60859 & 0,16206 & -1,05078 & 0,0365942 & -1,40626 & 0,528002 & -0,622101 & 0,977588 & 0,899172 \\
2,86413 & 2,76558 & 0,622101 & -5,04457 & -11,321 & -0,190942 & 1,23804 & -0,0431159 & 1,65688 & -0,622101 & 0,732971 & -1,15181 & -1,05942 \\
-4,50078 & -4,34591 & -0,977588 & 7,92717 & 17,7902 & 0,300052 & -1,9455 & 0,0677536 & -2,60367 & 0,977588 & -1,15181 & 1,80999 & 1,6648 \\
-4,13975 & -3,99731 & -0,899172 & 7,2913 & 16,3631 & 0,275983 & -1,78944 & 0,0623188 & -2,39482 & 0,899172 & -1,05942 & 1,6648 & 1,53126
\end{pmatrix}$$

Matica $\boldsymbol{E}$:

$$\begin{pmatrix}
29,4511 & -0,0923913 & 0,711957 & -25,788 & -81,1196 & -1,89674 & 2,2337 & 0,168478 & 3,16848 & -0,711957 & -0,36413 & -2,07065 & -2,71739 \\
-0,0923913 & 204,937 & -9,43297 & -38,7663 & -8,17029 & -0,193841 & -10,6141 & 0,762681 & -1,73732 & 0,432971 & 15,0344 & -2,91123 & 1,31159 \\
0,711957 & -9,43297 & 9,41486 & -0,918478 & -15,6486 & -0,780797 & 6,32065 & 0,436594 & -3,06341 & 0,585145 & -3,4221 & -2,19384 & -1,55797 \\
-25,788 & -38,7663 & -0,918478 & 65,5815 & 108,185 & 0,38587 & -2,17935 & -2,89674 & -1,89674 & 2,91848 & -8,25543 & 4,47283 & 4,1087 \\
-81,1196 & -8,17029 & -15,6486 & 108,185 & 9522,99 & 55,808 & -1,20652 & 24,1341 & 173,634 & 106,649 & -152,779 & 110,438 & 40,5797 \\
-1,89674 & -0,193841 & -0,780797 & 0,38587 & 55,808 & 15,3931 & -3,54891 & 2,58877 & -0,411232 & -0,219203 & -0,509058 & 1,87138 & 1,18116 \\
2,2337 & -10,6141 & 6,32065 & -2,17935 & -1,20652 & -3,54891 & 14,4946 & -0,527174 & 0,472826 & 0,679348 & -2,53804 & -6,94022 & -4,23913 \\
0,168478 & 0,762681 & 0,436594 & -2,89674 & 24,1341 & 2,58877 & -0,527174 & 11,1975 & 0,697464 & 1,56341 & 0,643116 & 0,132246 & 0,137681 \\
3,16848 & -1,73732 & -3,06341 & -1,89674 & 173,634 & -0,411232 & 0,472826 & 0,697464 & 29,6975 & 4,06341 & -0,356884 & 0,632246 & -0,862319 \\
-0,711957 & 0,432971 & 0,585145 & 2,91848 & 106,649 & -0,219203 & 0,679348 & 1,56341 & 4,06341 & 9,41486 & -11,5779 & -0,806159 & 0,557971 \\
-0,36413 & 15,0344 & -3,4221 & -8,25543 & -152,779 & -0,509058 & -2,53804 & 0,643116 & -0,356884 & -11,5779 & 65,567 & 1,75181 & 4,65942 \\
-2,07065 & -2,91123 & -2,19384 & 4,47283 & 110,438 & 1,87138 & -6,94022 & 0,132246 & 0,632246 & -0,806159 & 1,75181 & 15,6757 & 3,96377 \\
-2,71739 & 1,31159 & -1,55797 & 4,1087 & 40,5797 & 1,18116 & -4,23913 & 0,137681 & -0,862319 & 0,557971 & 4,65942 & 3,96377 & 14,8116
\end{pmatrix}$$

Matica $\boldsymbol{H}$:

$$\begin{pmatrix}
5,42689 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 14,3156 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 3,06836 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 8,09824 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 97,5858 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 3,92341 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 3,80717 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 3,34626 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5,44954 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3,06836 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8,09735 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3,95926 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3,848586
\end{pmatrix}$$

Príloha C.5: Kanonická diskriminančná analýza pre redukovanú dátovú maticu

Výberová kovariančná matica, kde `ProfitF`= 1.

$$\begin{pmatrix} 0,5295 & 0,0841 & 0,0242 & -0,5758 & 0,058 & 0,0899 & -0,0019 & -0,0377 & -0,0715 \\ 0,0841 & 3,1329 & -0,2198 & -0,4754 & -0,3275 & 0,1179 & 0,0643 & 0,014 & 0,0995 \\ 0,0242 & -0,2198 & 0,099 & -0,001 & 0,0821 & -0,0338 & -0,0101 & -0,0256 & -0,0087 \\ -0,5758 & -0,4754 & -0,001 & 1,1324 & -0,0068 & -0,1005 & 0,0454 & 0,0522 & 0,0802 \\ 0,058 & -0,3275 & 0,0821 & -0,0068 & 0,1971 & 0,0077 & 0,0068 & -0,1014 & -0,0831 \\ 0,0899 & 0,1179 & -0,0338 & -0,1005 & 0,0077 & 0,3053 & 0,0338 & 0,0039 & -0,0155 \\ -0,0019 & 0,0643 & -0,0101 & 0,0454 & 0,0068 & 0,0338 & 0,099 & 0,0034 & 0,0087 \\ -0,0377 & 0,014 & -0,0256 & 0,0522 & -0,1014 & 0,0039 & 0,0034 & 0,2382 & 0,0696 \\ -0,0715 & 0,0995 & -0,0087 & 0,0802 & -0,0831 & -0,0155 & 0,0087 & 0,0696 & 0,2551 \end{pmatrix}$$

Výberová kovariančná matica, kde `ProfitF`= 0.

$$\begin{pmatrix} 0,2446 & -0,1685 & -0,0163 & 0,0054 & -0,0163 & -0,038 & -0,0272 & -0,0163 & 0,0217 \\ -0,1685 & 2,7808 & 0,0199 & -0,7554 & 0,1793 & -0,3062 & -0,1069 & -0,154 & -0,1377 \\ -0,0163 & 0,0199 & 0,2156 & -0,038 & 0,1141 & -0,067 & 0,0453 & -0,0453 & -0,0507 \\ 0,0054 & -0,7554 & -0,038 & 0,6359 & -0,0815 & 0,1141 & 0,038 & 0,0924 & 0,0217 \\ -0,0163 & 0,1793 & 0,1141 & -0,0815 & 0,2446 & 0,0054 & 0,0163 & -0,1033 & -0,0217 \\ -0,038 & -0,3062 & -0,067 & 0,1141 & 0,0054 & 0,6938 & 0,1105 & 0,0199 & -0,0072 \\ -0,0272 & -0,1069 & 0,0453 & 0,038 & 0,0163 & 0,1105 & 0,2156 & -0,0417 & 0,0072 \\ -0,0163 & -0,154 & -0,0453 & 0,0924 & -0,1033 & 0,0199 & -0,0417 & 0,2156 & 0,0362 \\ 0,0217 & -0,1377 & -0,0507 & 0,0217 & -0,0217 & -0,0072 & 0,0072 & 0,0362 & 0,1449 \end{pmatrix}$$

Matica $\boldsymbol{B}$:

$$\begin{pmatrix} 11,1918 & 10,8067 & 2,4309 & -19,712 & 4,83773 & 6,47438 & -2,4309 & -4,50078 & -4,13975 \\ 10,8067 & 10,4348 & 2,34726 & -19,0337 & 4,67127 & 6,2516 & -2,34726 & -4,34591 & -3,99731 \\ 2,4309 & 2,34726 & 0,528002 & -4,28152 & 1,05078 & 1,40626 & -0,528002 & -0,977588 & -0,899172 \\ -19,712 & -19,0337 & -4,28152 & 34,7185 & -8,52065 & -11,4033 & 4,28152 & 7,92717 & 7,2913 \\ 4,83773 & 4,67127 & 1,05078 & -8,52065 & 2,09115 & 2,7986 & -1,05078 & -1,9455 & -1,78944 \\ 6,47438 & 6,2516 & 1,40626 & -11,4033 & 2,7986 & 3,74539 & -1,40626 & -2,60367 & -2,39482 \\ -2,4309 & -2,34726 & -0,528002 & 4,28152 & -1,05078 & -1,40626 & 0,528002 & 0,977588 & 0,899172 \\ -4,50078 & -4,34591 & -0,977588 & 7,92717 & -1,9455 & -2,60367 & 0,977588 & 1,80999 & 1,6648 \\ -4,13975 & -3,99731 & -0,899172 & 7,2913 & -1,78944 & -2,39482 & 0,899172 & 1,6648 & 1,53126 \end{pmatrix}$$

Matica $\boldsymbol{E}$:

$$\begin{pmatrix} 29,4511 & -0,0923913 & 0,711957 & -25,788 & 2,2337 & 3,16848 & -0,711957 & -2,07065 & -2,71739 \\ -0,0923913 & 204,937 & -9,43297 & -38,7663 & -10,6141 & -1,73732 & 0,432971 & -2,91123 & 1,31159 \\ 0,711957 & -9,43297 & 9,41486 & -0,918478 & 6,32065 & -3,06341 & 0,585145 & -2,19384 & -1,55797 \\ -25,788 & -38,7663 & -0,918478 & 65,5815 & -2,17935 & -1,89674 & 2,91848 & 4,47283 & 4,1087 \\ 2,2337 & -10,6141 & 6,32065 & -2,17935 & 14,4946 & 0,472826 & 0,679348 & -6,94022 & -4,23913 \\ 3,16848 & -1,73732 & -3,06341 & -1,89674 & 0,472826 & 29,6975 & 4,06341 & 0,632246 & -0,862319 \\ -0,711957 & 0,432971 & 0,585145 & 2,91848 & 0,679348 & 4,06341 & 9,41486 & -0,806159 & 0,557971 \\ -2,07065 & -2,91123 & -2,19384 & 4,47283 & -6,94022 & 0,632246 & -0,806159 & 15,6757 & 3,96377 \\ -2,71739 & 1,31159 & -1,55797 & 4,1087 & -4,23913 & -0,862319 & 0,557971 & 3,96377 & 14,811 \end{pmatrix}$$

Matica $\boldsymbol{H}$:

$$\begin{pmatrix} 5,42689 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 14,3156 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3,06836 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8,09824 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3,80717 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5,44954 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3,06836 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3,95926 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3,84858 \end{pmatrix}$$

Príloha C.6: Matica vzdialeností a matica kvadratických diskriminačných funkcií

Matica vzdialeností

$$\begin{pmatrix}
3,34053 & 14,546 \\
2,61353 & 12,9805 \\
0,884458 & 8,56538 \\
0,285477 & 6,35297 \\
0,281779 & 6,33548 \\
1,75249 \times 10^{-6} & 3,93977 \\
0,57771 & 7,54207 \\
0,930547 & 1,04359 \\
2,29585 & 0,221846 \\
3,31683 & 0,0272227 \\
2,96664 & 13,7538 \\
6,42932 \times 10^{-8} & 3,94604 \\
0,194002 & 2,38936 \\
0,187886 & 5,85479 \\
1,42044 & 10,0999 \\
0,190178 & 5,86756 \\
1,4703 & 0,598539 \\
0,120398 & 2,68706 \\
0,278117 & 2,12822 \\
0,953115 & 8,77632 \\
0,755825 & 8,15441 \\
0,421386 & 6,94508 \\
0,0101068 & 3,55578 \\
1,27348 & 0,735692 \\
0,26956 & 6,27704 \\
1,97809 & 11,5101 \\
1,40605 & 10,0614 \\
0,719492 & 1,295 \\
3,34053 & 14,546 \\
0,974859 & 0,997722 \\
1,11173 & 0,868294 \\
2,30889 & 0,217817 \\
1,94994 & 11,4421 \\
0,00703718 & 4,28531 \\
4,42416 & 0,0137259 \\
1,67932 & 0,476549 \\
0,0110382 & 3,53872 \\
3,40747 & 0,0196777 \\
0,163432 & 5,71438 \\
0,579603 & 1,50036 \\
0,0422219 & 4,8035 \\
0,40029 & 1,83203 \\
0,423436 & 6,9534 \\
0,983826 & 0,988693 \\
0,0545182 & 4,92707 \\
0,0391323 & 3,19834 \\
0,985891 & 0,986625 \\
1,13191 & 0,850632 \\
0,597035 & 1,47265 \\
10,1695 & 1,44664 \\
1,27447 & 0,734941 \\
11,1431 & 1,82769 \\
1,39682 & 0,646962 \\
2,93063 & 0,075242 \\
0,423917 & 1,78255 \\
6,96726 & 0,426862 \\
0,719492 & 1,295 \\
6,40436 & 0,296452 \\
7,72344 & 0,628688 \\
2,28671 & 0,224698 \\
9,33587 & 1,14331 \\
4,74133 & 0,0365763 \\
3,52913 & 0,0115804 \\
3,02426 & 0,0610934 \\
7,91494 & 0,684161 \\
1,64771 & 0,493617 \\
7,93202 & 0,68919 \\
5,57276 & 0,14022 \\
8,70541 & 0,929841 \\
5,07951 & 0,0715916
\end{pmatrix}$$

Matica kv. diskriminačných funkcií

$$\begin{pmatrix}
3,44347 & -12,3463 \\
2,51706 & -10,9129 \\
1,19811 & -8,63459 \\
-0,870833 & -16,9943 \\
0,597705 & -11,076 \\
-0,838924 & -7,07971 \\
2,63307 & -2,85094 \\
0,016601 & -1,03184 \\
-3,63836 & -3,63527 \\
-0,854569 & -0,150882 \\
4,04895 & -12,5124 \\
2,03724 & -4,00326 \\
0,0846222 & -2,53225 \\
-1,89866 & -9,81336 \\
1,70699 & -8,69976 \\
2,08535 & -6,21971 \\
2,15666 & 0,890686 \\
3,77385 & -0,512874 \\
2,00292 & -0,309221 \\
1,57901 & -13,8016 \\
3,12239 & -6,85789 \\
1,88762 & -6,62278 \\
2,73495 & -3,62317 \\
-0,2428 & -2,74874 \\
0,826774 & -11,8925 \\
3,00537 & -16,9273 \\
4,08612 & -8,67937 \\
1,31631 & 0,626745 \\
3,44347 & -12,3463 \\
1,51901 & -7,20539 \\
1,05093 & -1,43057 \\
1,64258 & -3,39758 \\
2,77547 & -9,05775 \\
1,96167 & -0,800345 \\
-2,28302 & 0,221509 \\
-4,4269 & -1,23859 \\
1,8175 & -3,56222 \\
0,385084 & -1,72923 \\
3,09423 & -8,76077 \\
-1,7621 & -7,71916 \\
-0,536351 & -11,6417 \\
2,19269 & -0,599615 \\
2,3772 & -14,3016 \\
0,313902 & -3,18996 \\
1,2572 & -6,02628 \\
2,34266 & -2,08206 \\
0,350104 & -0,665565 \\
-3,1751 & -1,41636 \\
0,559808 & 0,544838 \\
-9,50873 & -0,69291 \\
-1,64305 & 0,785175 \\
-4,55992 & -0,675813 \\
-2,59667 & -0,174659 \\
-2,81065 & 1,93081 \\
-0,165423 & -2,90551 \\
-4,44338 & -0,612013 \\
1,31631 & 0,626745 \\
-3,96306 & 1,64195 \\
-5,4325 & 1,17575 \\
-8,3289 & -1,91422 \\
-8,30715 & -0,776021 \\
-3,91837 & 0,897195 \\
-5,04962 & 0,76408 \\
0,556815 & 2,05877 \\
-2,54896 & 2,46083 \\
-2,27061 & 0,352583 \\
-1,78716 & 1,05455 \\
-3,24699 & -0,317989 \\
-13,8735 & -2,5294 \\
-3,13174 & -0,280326
\end{pmatrix}$$

Príloha C.7: Matica vzdialeností a matica kvadratických diskriminačných funkcií nových žiadostí

Matica vzdialeností

$$\begin{pmatrix} 0,0181642 & 3,42781 \\ 0,0698395 & 5,06467 \\ 11,3184 & 1,8991 \\ 0,0506356 & 4,88956 \\ 0,00747377 & 3,60909 \\ 0,926552 & 8,69534 \\ 0,768398 & 1,23127 \\ 3,04248 & 0,0585342 \\ 0,0000266813 & 3,96558 \\ 4,36437 & 0,0105878 \\ 0,911268 & 1,06421 \\ 3,97413 & 0,0000534542 \end{pmatrix}$$

Matica kv. diskriminačných funkcií

$$\begin{pmatrix} 0,599887 & -5,19989 \\ -3,98662 & -8,27684 \\ -16,1956 & -7,80866 \\ 1,99992 & -2,7989 \\ -2,33153 & -5,09275 \\ -0,969286 & -16,1033 \\ -10,7986 & -15,5914 \\ -4,75287 & -2,39605 \\ -6,60846 & -5,8397 \\ -7,44539 & -3,1209 \\ -3,18793 & -4,31795 \\ -6,59628 & -2,84671 \end{pmatrix}$$

Príloha D.1: Kódovanie premenných na dummy

Stvrt	Dummy	
Kategória	$Stvrt_1$	$Stvrt_2$
1	0	0
2	1	0
3	0	1

Platenie	Dummy	
Kategória	$Platenie_1$	$Platenie_2$
1	0	0
2	1	0
3	0	1

DalsieP	Dummy		
Kategória	$DalsieP_1$	$DalsieP_2$	$DalsieP_3$
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

PC	Dummy			
Kategória	PC_1	PC_2	PC_3	PC_4
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

Mesto	Dummy				
Kategória	$Mesto_1$	$Mesto_2$	$Mesto_3$	$Mesto_4$	$Mesto_5$
1	0	0	0	0	0
2	1	0	0	0	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	1	0
6	0	0	0	0	1

Premenné s dvoma kategóriami t.j. `Adr, TO, Pohlavie, SP, BL, Zlavy1 a Zlavy2` sme kódovali 1 pre 1. kategóriu a 0 pre 2. kategóriu.

Príloha D.2: Model logistickej regresie

Premenná	Odhad koeficientu	Štandardná chyba	Waldova štatistika	p-hodnota
$Kontanta$	6,84639	2,66401	2,570	0,01017
Adr	2,82889	1,98807	1,423	0,15475
$Mesto_1$	-5,03134	0,91834	-5,479	0,00000
$Mesto_3$	-3,01892	0,93750	-3,220	0,00128
$Mesto_4$	-4,29902	0,84382	-5,095	0,00000
$Mesto_5$	-2,11106	1,29099	-1,635	0,10200
$Platenie_1$	-1,23612	1,32880	-0,930	0,35224
PC_3	5,27350	0,69333	7,606	0,00000
PC_4	4,78735	0,04319	110,852	0,00000
$Stvrt_1$	-5,13568	1,83707	-2,796	0,00518
$Stvrt_2$	-7,86039	1,92852	-4,076	0,00005